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

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

Abstract—With the unprecedented developments in deep learning, automatic segmentation of main abdominal organs seems to be a solved problem as state-of-the-art (SOTA) methods have achieved comparable results with inter-rater variability on many benchmark datasets. However, most of the existing abdominal datasets only contain single-center, single-phase, single-vendor, or single-disease cases, and it is unclear whether the excellent performance can generalize on diverse datasets. This paper presents a large and diverse abdominal CT organ segmentation dataset, termed AbdomenCT-1K, with more than 1000 (1K) CT scans from 12 medical centers, including multi-phase, multi-vendor, and multi-disease cases. Furthermore, we conduct a large-scale study for liver, kidney, spleen, and pancreas segmentation and reveal the unsolved segmentation problems of the SOTA methods, such as the limited generalization ability on distinct medical centers, phases, and unseen diseases. To advance the unsolved problems, we further 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 AbdomenCT-1K dataset will promote future in-depth research towards clinical applicable abdominal organ segmentation methods.

1 Introduction

Abdominal 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], [4]. In particular, computed tomography (CT) scan is one of the most commonly used modalities for the abdominal 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, the following reasons make organ segmentation a difficult task. First, the contrast of soft tissues is usually low. Second, organs may have complex morphological structures and heterogeneous lesions. Last but not least, different scanners and CT phases can lead to significant variances in organ appearances. Figure 1 presents some examples of these challenging situations.

Manual contour delineation of target organs is labor-intensive and time-consuming, and also suffers from inter- and intra- observer variability [5]. Therefore, automatic segmentation methods are highly desired in clinical studies. In the past two decades, many abdominal segmentation methods have been proposed and massive progress has been achieved continuously in the era of deep learning. For instance, from a recently presented review work, liver segmentation can reach an accuracy of 95% in terms of Dice similarity coefficient (DSC) [6]. In a recent work for spleen segmentation [7], 96.2% DSC score was reported. However, most of the existing abdominal datasets only contain single-center, single-phase, single-vendor, and single-disease cases, which makes it unclear that if the performance obtained on these datasets can generalize well on more diverse datasets. Therefore, it is worth re-thinking that *is abdominal organ segmentation a solved problem?*

To answer this question, in this paper, we first build

- *This project is supported by China's Ministry of Science and Technology (No. 2020YFA0713800) and National Natural Science Foundation of China (No. 11971229, No. 12090023). 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. This work is done when Yao Zhang is an intern at AI Lab., Lenovo Research.*
- *Song Gu is with School of Automation, Nanjing University of Information Science and Technology, P.R. China.*
- *Cheng Zhu is with Shenzhen Haichuang Medical CO., LTD., P.R. China.*
- *Cheng Ge is with Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, P.R. China.*
- *Yichi Zhang is with School of Biological Science and Medical Engineering, Beihang University, China.*
- *Xingle An is with Beijing Infervision Technology CO. LTD., P.R. China.*
- *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.*
- *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.*
- *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 Computational Biology, University of Southern California, US.*
- *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.*

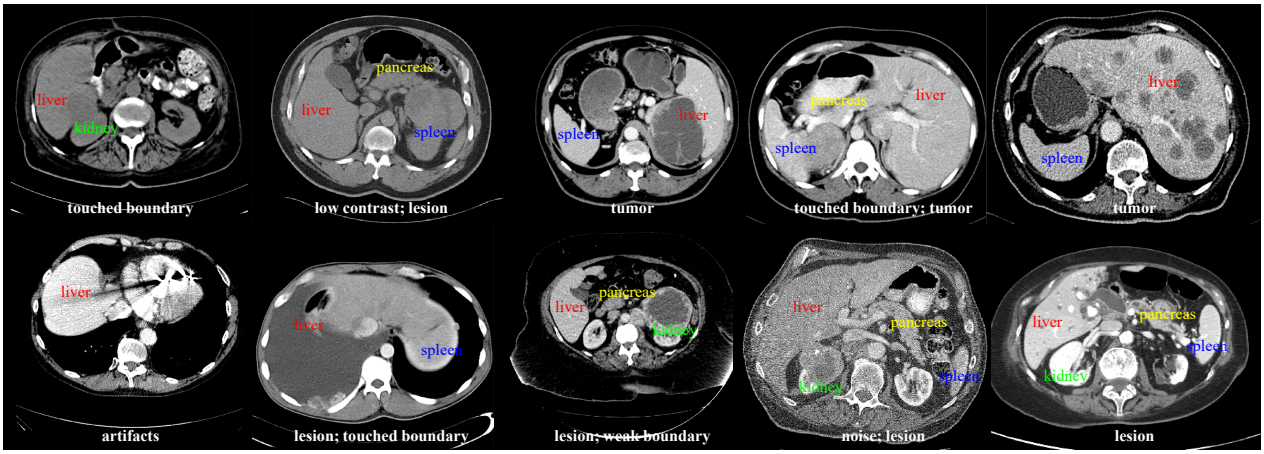


Fig. 1: Examples of abdominal organs in CT scans, including multi-center, multi-phase, multi-vendor, and multi-disease cases.

a large and diverse abdominal CT organ segmentation dataset, namely AbdomenCT-1K. Then, we investigate the current limitations of the existing solutions based on the dataset. Finally, we provide four elaborately designed benchmarks for the challenging and practical problems of abdominal organ segmentation. In the following subsections, we will summarize the limitations of the existing methods and benchmarks, and then we will briefly present the contributions of our work.

1.1 Limitations of existing abdominal organ segmentation methods and benchmark datasets

A clinically feasible segmentation algorithm should not only reach high accuracy, but also can generalize well on data from different sources [8], [9]. However, despite the encouraging progress of deep learning-based approaches and benchmarks, the methods and benchmarks still have some limitations that are briefly summarized as follows.

1) **Lack of a large-scale and diverse dataset.** Evaluating the generalization ability on a large-scale and diverse dataset is highly demanded, but there exist no such kind of public dataset. As shown in Table 1, most of the existing benchmark datasets either have a small number of cases or are collected 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 are trained and evaluated on small publicly available datasets. It is unclear whether the proposed methods can generalize well on other testing cases, especially when the testing set 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 [10], [11], [12], [13], because it is labor-intensive and time-consuming to obtain manual annotations. The availability of benchmarks plays an important role in the progress of methodology developments. For example, the SOTA performance of video segmentation has been considerably improved by the DAVIS video object segmentation benchmarks [14], including semi-supervised, interactive and unsupervised tasks [15]. However, no such kind of benchmark exists 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 metrics.** Many of the existing benchmarks [16], [17] only use the region-based measurement (i.e., DSC) to rank segmentation methods. Boundary accuracy is also important in clinical practice [18], [19], but it is insufficient to measure the boundary accuracy by DSC as demonstrated and analyzed in Figure 5.

1.2 Contributions

To address the above limitations, in this work, we firstly create a large-scale abdominal multi-organ CT dataset by extending the existing benchmark datasets with more organ annotations, including LiTS [16], MSD [20], KiTS [17],NIH-Pancreas [21], [22], [23]. Specifically, our dataset, termed AbdomenCT-1K, includes 1112 CT scans from 12 medical centers with multi-center, multi-phase, multi-vendor, and multi-disease cases. We annotate the liver, kidney, spleen, and pancreas for all cases. Figure 3 and Table 1 illustrate the proposed AbdomenCT-1K dataset and list the main different points between our dataset and the existing abdominal organ datasets. 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 [24]) on the AbdomenCT-1K dataset for single organ and multi-organ segmentation tasks. In addition to the widely used DSC, we add the normalized surface Dice (NSD) [25] as a boundary-based

evaluation metric because the segmentation accuracy in organ boundaries is also very important in clinical practice [18], [19]. Based on the results, we find that the answer is **Yes** for some ideal or easy situations, but abdominal organ segmentation is still an unsolved problem in the challenging situations, especially in the authentic clinical practice, e.g., the testing set is from a new medical center and/or contains 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 Section 4. 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, which are increasingly drawing attention in the medical image analysis community. Figure 2 presents an overview of our new abdominal organ benchmarks.

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 1112 CT scans from 12 medical centers including multi-phase, multi-vendor, and multi-disease cases. The annotations include 4446 organs (liver, kidney, spleen, and pancreas) that are significantly larger than existing abdominal organ segmentation datasets. More importantly, our dataset provides a platform for researchers to pay more attention to the generalization ability of the algorithms when developing new segmentation methodologies, which is critical for the methods to be applied in clinical practice.
2) We conduct a large-scale study for liver, kidney, spleen, and pancreas segmentation based on the AbdomenCT-1K dataset and the SOTA method nnU-Net [24]. The extensive experiments identify some solved problems and, more importantly, reveal the unsolved problems in abdominal organ segmentation.
3) We establish, for the first time, four new abdominal multi-organ segmentation benchmarks for fully supervised[1], semi-supervised[2], weakly supervised[3], and continual learning[4]. These benchmarks can provide a standardized and fair evaluation of abdominal organ segmentation methods. Moreover, we also develop and provide out-of-the-box baseline solutions with the SOTA method for each task. Our dataset, code, and trained models are publicly available at https://github.com/JunMa11/AbdomenCT-1K.

1. https://abdomenct-1k-fully-supervised-learning.grand-challenge.org/
2. https://abdomenct-1k-semi-supervised-learning.grand-challenge.org/
3. https://abdomenct-1k-weaklysupervisedlearning.grand-challenge.org/
4. https://abdomenct-1k-continual-learning.grand-challenge.org/

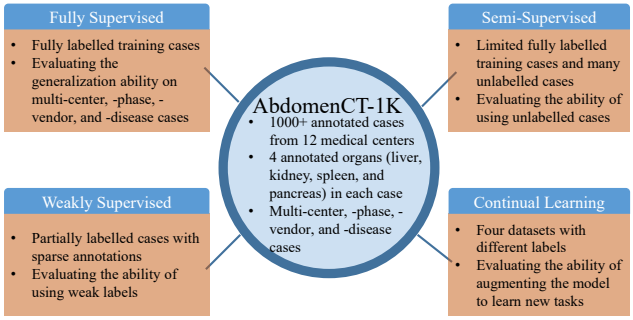


Fig. 2: Task overview and the associated features.

Abdominal organ segmentation in CT scans is one of the most popular segmentation tasks and there are more than 4000 teams[5] working on existing benchmarks. We believe that our AbdomenCT-1K and carefully designed benchmarks can again attract the attention of the community to focus on the more challenging and practical problems in abdominal organ segmentation.

The rest of the paper is organized as follows. First, in Section 2, the related work, including a review of abdominal organ segmentation methods and existing datasets, is presented. Then, in Section 3, we describe the created AbdomenCT-1K dataset. Afterwards, we conduct a comprehensive study for abdominal organ segmentation with the SOTA method nnU-Net [24] in Section 4, where the solved and unsolved problems for abdominal organ segmentation are also presented. Next, in order to address these unsolved problems, we set up four new benchmarks in Section 5, including fully supervised, semi-supervised, weakly supervised, and continual learning of abdominal organ segmentation, respectively. Finally, in Section 6, the conclusions are drawn.

2 Related Work

2.1 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.

Model-based methods usually formulate the image segmentation as an energy functional minimization problem or explicitly match a shape template or atlas to a new image, such as variational models [26], statistical shape models [27], and atlas-based methods [28]. Level set methods or active contour models are one of the most popular variational models. They provide a natural way to drive the curves to delineate the structure of interest [29], [30], [31]. Different from the level set methods, statistical shape models, such as the well-known active shape model, represent the shape of an object by a set of boundary points that are constrained by the point distribution model. Then, the model iteratively deforms the points to fit to the object in a new image [1], [32]. Atlas-based methods usually construct one or multiple organ atlas with annotated cases. Then, label fusion is used to propagate the atlas annotations to a target image via

5. https://grand-challenge.org/challenges/

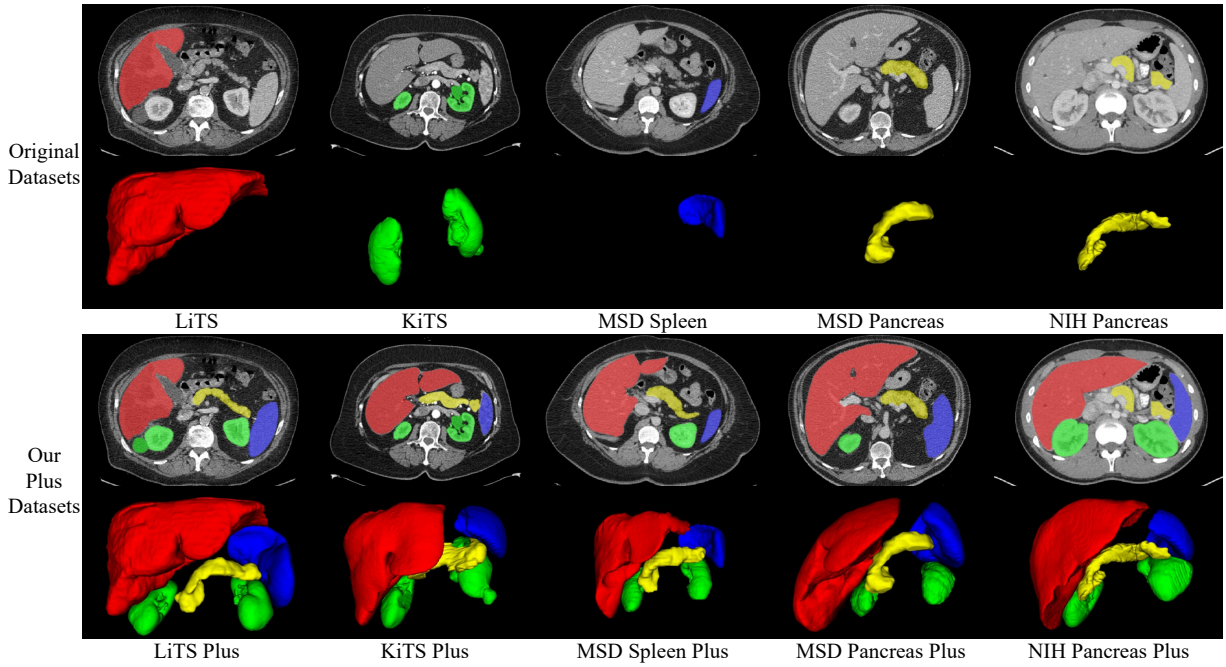


Fig. 3: Overview of the existing abdominal CT datasets and our augmented (plus) abdominal datasets. Red, green, blue, and yellow regions denote liver, kidney, spleen, and pancreas, respectively.

registration between the atlas image and the target image [33], [34], [35]. Although these model-based methods have transparent principles and well-defined formulations, they usually fail 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 discriminative features from annotated CT scans to distinguish target organs and other tissues. Since 2015, deep convolutional neural network (CNN)-based methods [36], which neither rely on hand crafted features nor rely on anatomical correspondences, have been successfully introduced into abdominal organ segmentation and reach SOTA performances [37], [38]. These approaches can be briefly classified into well-known supervised learning methods and recently emerging annotation-efficient deep learning approaches. In the following paragraphs, we will introduce the two categories respectively.

One group of the supervised organ segmentation methods is single organ segmentation. For example, Seo *et al.* proposed a modified U-Net [39] to segment liver and liver tumors. In [40], a shape-aware method, which incorporated prior knowledge of the target organ shape into a CNN backbone, was proposed and achieved encouraging performance on liver segmentation task. While U-Net is a welcomed network structure, other backbone designs are also proposed for abdominal organ segmentation, such as progressive holistically-nested network (PHNN) [41], [42] and progressive semantically-nested networks (PSNNs) [43]. In [44], CNNs were employed to segment the pancreas. Pancreas segmentation was treated as a more challenging task compared to the liver and the kidney segmentation. Therefore, two-stage cascaded approaches were proposed [45], [46], [47], where pancreas was located first, then a new network was employed to refine the segmentation. Moreover, in [48], a level set regression network was developed to obtain more accurate segmentation in pancreas boundaries. Instead of designing network structures empirically, Neural Architecture Search (NAS) technique was also introduced into organ segmentation [49], [50], [51] by designing efficient differentiable neural architecture search strategies.

The other group of the supervised organ segmentation methods is multi-organ segmentation [4], [12], [52], [53], where multiple organs are segmented simultaneously. Fully convolutional networks (FCN)-based methods have been widely applied to multi-organ segmentation. Early works include applying FCN alone [53], [54] and the combinations of FCN with pre- or/and post-processing [55], [56]. However, compared to the single organ segmentation task, multi-organ segmentation is more challenging. As shown in Figure 1, the weak boundaries between organs on CT scans and the variations of the size of different organs, make the multi-organ segmentation task harder [4]. In order to address the difficulties, cascaded networks were employed to organ segmentation. In [4], a two-stage segmentation method was proposed. An organ segmentation probability map was first computed in the first stage and was combined with 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. Other similar strategies were proposed [52], [57], where the first stage networks played different roles. For example, in [52], a candidate region was generated and sent to the second stage. In [57], low resolution segmentation maps were extracted from the first stage. Moreover, in [58], 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 [59] was proposed and showed improved performance. In order to reduce the choices of the number of architecture layers, kernel sizes, *etc.*, in [60], trainable 3D convolutional kernel with learnable filter coefficients and spatial offsets was presented and show its benefits to capture large spatial context as well as the design of networks. Noticeably, in [24], nnU-Net, a U-Net [36]-based segmentation framework, was proposed and achieved state-of-the-art performances on both single organ and multi-organ segmentation tasks, including liver, kidney, pancreas, and spleen.

Recently, **annotation-efficient methods**, such as semi-supervised learning, weakly supervised learning, and continual learning, have received great attention in both computer vision and medical image analysis communities [6], [10], [11]. 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 partially labelled organs.

Semi-supervised learning aims to combine a small amount of labelled data with a large amount of unlabelled data, which is an effective way to explore knowledge from the unlabelled data. It is a promising and active research direction in machine learning [61] as well as medical image analysis [10]. Among the semi-supervised approaches, pseudo label-based methods are regarded as simple and efficient solutions [62], [63]. In [64], 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 unlabelled dataset were computed by the trained model. Finally, a student model was trained on the combination of both the labelled and unlabelled data. Besides, other strategies are also explored. For example, in [65], in addition to the Dice loss computed from labelled data, a quality assurance-based discriminator module was proposed to supervise the learning on the unlabelled data. In [66], a co-training strategy was proposed to explore unlabelled data. The proposed framework, trained on a small single phase dataset, can adapt to unlabelled multi-center and multi-phase clinical data. Moreover, an uncertainty-aware multi-view co-training (UMCT) approach was proposed in [67], which achieves superior performance on multi-organ and pancreas datasets.

Weakly supervised learning is to explore the use of weak annotations, such as slice-level annotations, sparse annotations, and noisy annotations [11]. For organ segmentation, in [68], a classification forest-based weakly supervised organ segmentation method was proposed for livers, spleens and kidneys, where the labels are scribbles on organs. Besides, image-level labels-based pancreas segmentation was explored in [69]. Although there are limited studies 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 [70], points [71], [72], scribbles [73], [74], image-level labels [75], [76], [77].

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 [78], [79], [80]. The learned knowledge of a model can be interfered with the new information which we train the model with. As a result, the performance of the old task could decrease. Therefore, continual learning has attracted growing attention in the past years [81], such as object recognition [82], [83] and classification [84]. Besides, tailored datasets and benchmarks for continual learning have been also proposed in the computer vision community, e.g. the object recognition dataset and benchmark CORe50 [82], iCubWorld datasets[6], and the CVPR2020 CLVision challange[7]. However, to the best of our knowledge, there is no continual learning work for abdominal organ segmentation. Therefore, applying this new emerging technique to tackle organ segmentation tasks is still in demand.

2.2 Existing abdominal CT organ segmentation benchmark datasets

In addition to the promising progress in abdominal organ segmentation methodologies, segmentation benchmark datasets are also evolved, where the datasets contain 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) [85] benchmark dataset consists of 50 abdominal CT scans 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 labelled by two experienced undergraduate students, and verified by a radiologist.

NIH Pancreas dataset [21], [22], [23], from US National Institutes of Health (NIH) Clinical Center, consists of 80 abdominal contrast enhanced 3D CT images. The CT scans have resolutions of 512×512 pixels with varying pixel sizes and slice thickness between $1.5-2.5$ mm. Among these cases, 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

6. https://robotology.github.io/iCubWorld/#publications
7. https://sites.google.com/view/clvision2020/challenge
8. https://www.synapse.org/#!Synapse:syn3193805/wiki/89480
9. https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT
10. http://www.visceral.eu/benchmarks/
11. https://competitions.codalab.org/competitions/15595
12. http://medicaldecathlon.com/
13. http://medicaldecathlon.com/
14. https://zenodo.org/record/1169361#.YMRb9NUza70
15. https://chaos.grand-challenge.org/
16. https://kits19.grand-challenge.org/
17. https://wiki.cancerimagingarchive.net/display/Public/CT-ORG%3A+CT+volumes+with+multiple+organ+segmentations

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

Dataset Name (abbr.)	Target	# of Tr/Ts	# of Centers	Source and Year
Multi-atlas Labelling Beyond the Cranial Vault (BTCV) [8] [85]	13 organs	30/20	1	MICCAI 2015
NIH Pancreas [9] [21], [22], [23]	Pancreas	80	1	The Cancer Imaging Archive 2015
VISCERAL Anatomy Benchmark [10] [86]	20 anatomical structures	80/40	1	ISBI and ECIR 2015
Liver Tumor Segmentation Benchmark (LiTS) [11] [16]	Liver and tumor	131/70	7	ISBI and MICCAI 2017
Medical Segmentation Decathlon (MSD) Pancreas [12] [20]	Pancreas and tumor	281/139	1	MICCAI 2018
Medical Segmentation Decathlon (MSD) Spleen [13] [20]	Spleen	41/20	1	MICCAI 2018
Multi-organ Abdominal CT Reference Standard Segmentation [14] [53]	8 organs	90	2	Zenodo 2018
Combined Healthy Abdominal Organ Segmentation (CHAOS) [15] [87]	Liver	20/20	1	ISBI 2019
Kidney Tumor Segmentation Benchmark (KiTS) [16] [88]	Kidney and tumor	210/90	1	MICCAI 2019
CT-ORG [17] [89]	liver, lungs, bladder, kidney, bones and brain	119/21	8	The Cancer Imaging Archive 2020
AbdomenCT-1K (ours)	Liver, kidney, spleen, and pancreas	**1112**	**12**	2021

pancreas was manually labelled slice-by-slice by a medical student and then verified/modified by an experienced radiologist.

VISCERAL Anatomy Benchmark [86] consists of 120 CT and MR patient volumes.Volumes from 4 different imaging modalities and field-of-views compose the training set. Each group contains 20 volumes, which adds up to 80 volumes in the training set. In each volume, 20 abdominal structures were manually annotated to build a standard Gold Corpus containing a total of 1295 structures and 1760 landmarks.

LiTS (Liver Tumor Segmentation) dataset [16] 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 collected from 7 medical centers and the corresponding patients have 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 [20] consists of 281 training cases with pancreas and tumor annotations and 139 testing cases with hidden annotations. The dataset is 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 [20] 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.

Multi-organ Abdominal CT Reference Standard Segmentations [53] is composed of 90 abdominal CT images and corresponding reference standard segmentations of 8 organs. The CT images are from the Cancer Imaging Archive (TCIA) Pancreas-CT dataset with pancreas segmentations and the Beyond the Cranial Vault (BTCV) challenge with segmentations of all organs except duodenum. The unsegmented organs were manually labelled by an imaging research fellow under the supervision of a board-certified radiologist.

CHAOS (Combined Healthy Abdominal Organ Segmentation) dataset [87] 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 the healthy population.

KiTS (Kidney Tumor Segmentation) dataset [88] includes 210 training cases with kidney and kidney tumor annotations and 90 testing cases with hidden annotations, which are provided by the 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.

CT-ORG [89] is a diverse dataset of 140 CT images containing 6 organ classes, where 131 are dedicated CT and 9 are the CT component from PET-CT exams. These CT images are from 8 different medical centers. Patients were included based on the presence of lesions in one or more of the labelled organs. Most of the images exhibit liver lesions, both benign and malignant.

3 AbdomenCT-1K dataset

3.1 Dataset motivation and details

Most existing abdominal organ segmentation datasets 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, namely AbdomenCT-1K, consists of 1112 3D CT scans from five existing datasets: LiTS (201 cases) [16], KiTS (300 cases) [17], MSD Spleen (61 cases) and Pancreas (420 cases) [20], NIH Pancreas (80 cases) [21], [22], [23], and a new dataset from Nanjing University (50 cases). The 50 CT scans in the Nanjing University dataset are from 20 patients with pancreas cancer, 20 patients with colon cancer, and 10 patients with liver cancer. The number of plain phase, artery phase, and portal phase scans are 18, 18, and 14 respectively. The CT scans have resolutions of 512×512 pixels with varying pixel sizes and slice thicknesses between 1.25-5 mm, acquired on GE multi-detector spiral CT. 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 share or redistribute them in any format.

The original datasets only provide annotations of one single organ, while our dataset contains annotations of four organs for all cases in each dataset as shown in Figure 3. In order to distinguish from the original datasets, we term our multi-organ annotations as plus datasets (e.g., the multi-organ LiTS dataset is termed as LiTS Plus dataset in this paper). Figure 4 presents the organ volume and contrast phase distributions in AbdomenCT-1K. The other information (e.g., CT scanners, the distribution of the Hounsfield unit (HU) value, image size, and image spacing.) is presented in the supplementary (Supplementary Table 1).

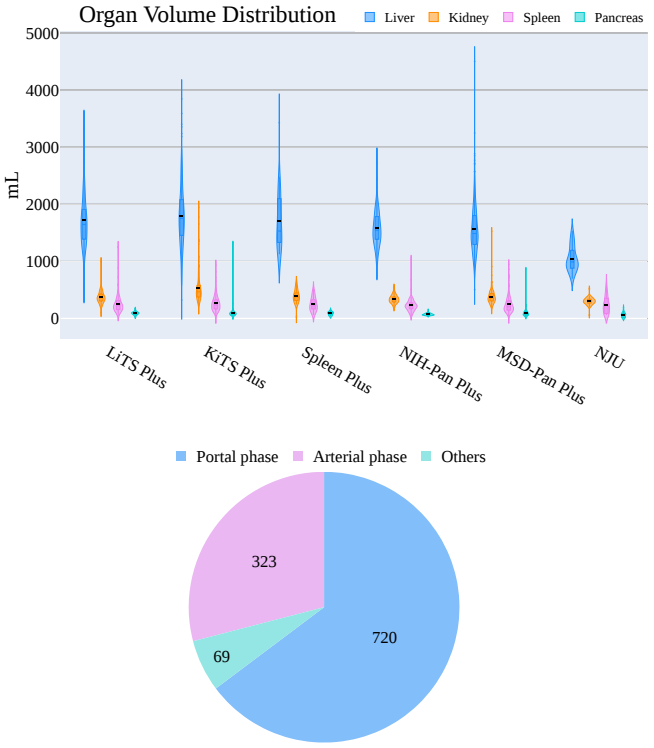


Fig. 4: Organ volume and contrast phase distributions in AbdomenCT-1K.

3.2 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. 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, one senior radiologist with more than 10-years experience verifies and refines the annotations. All the annotations are applied to axial images. To reduce inter-rater annotation variability, we introduce three hierarchical strategies to improve the label consistency. Specifically,

- before annotation, all raters are required to learn the existing organ annotation protocols, aiming to ensure that the annotation protocols are consistent in raters and the existing datasets;
- during annotation, the obvious label errors in existing datasets are fixed and all annotations are finally checked and revised by an experienced senior radiologist (10+ years specialized in the abdomen);
- after annotation, we train five-fold cross-validation U-Net models to find the possible segmentation errors. The cases with low DSC or NSD scores are double-checked by the senior radiologist.

In addition, we invite two radiologists to annotate the 50 cases in the Nanjing University dataset and present their inter-rater variability in Table 2.

TABLE 2: Quantitative analysis of inter-rater variability between two radiologists.

Organ	Liver	Kidney	Spleen	Pancreas
DSC (%)	98.4 ± 0.52	98.7 ± 0.53	98.6 ± 0.84	93.8 ± 7.78
NSD (%)	95.7 ± 3.04	98.7 ± 2.05	98.2 ± 4.18	92.5 ± 9.40

3.3 Backbone network

The legendary U-Net ([36], [90]) has been widely used in various medical image segmentation tasks, and many variants have been proposed to improve it. However, recent studies [17], [24] demonstrate 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) [24] 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 performances than most specialized deep learning pipelines in 19 public international segmentation competitions and set a new SOTA in 49 tasks. Currently, nnU-Net is still the SOTA method in many segmentation tasks [91]. Thus, we employ nnU-Net as our backbone network[8]. Specifically, the network input is configured with a batch size of 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,

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

scaling, adding Gaussian Noise, gamma correction. The loss function is a combination of Dice loss [92] and cross-entropy loss because compound loss functions have been proved to be robust in many segmentation tasks [93]. All the models are trained for 1000 epochs with the above hyper-parameters on NVIDIA TITAN V100 or 2080Ti GPUs.

3.4 Evaluation metrics

Motivated by the evaluation methods of the well-known medical image segmentation decathlon[9], we 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) [25], 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 scores in $[0, 1]$ and higher scores indicate 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:

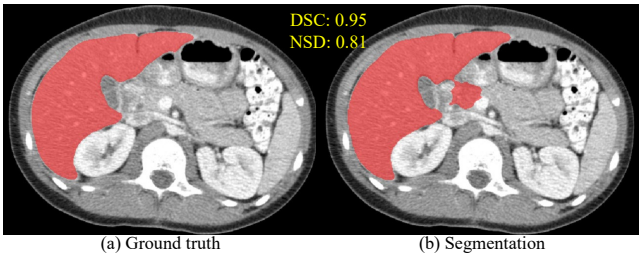
- Region-based measure: DSC

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

- Boundary-based measure: NSD

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

where $B_{\partial G}^{(\tau)} = \{x \in R^3 \,|\, \exists \tilde{x} \in \partial G, \, ||x - \tilde{x}|| \leq \tau\}$, $B_{\partial S}^{(\tau)} = \{x \in R^3 \,|\, \exists \tilde{x} \in \partial S, \, ||x - \tilde{x}|| \leq \tau\}$ denote the border region of the ground truth and the segmentation surface at tolerance τ, respectively. In this paper, we set the tolerance τ as 1mm.



(a) Ground truth (b) Segmentation

Fig. 5: 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 [16], [17], while NSD can provide additional complementary information to the segmentation quality. Figure 5 presents a liver segmentation example to illustrate the features of NSD. An obvious segmentation error can be found on the right side boundary of the liver. However, the DSC score is still very high that cannot well reflect the boundary error, while NSD is sensitive to this boundary error and thus a low score is obtained. In many clinical tasks, such as preoperative planning and organ transplant, boundary errors are critical [18], [19] 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.

9. http://medicaldecathlon.com/

4 A LARGE-SCALE STUDY ON FULLY SUPERVISED ORGAN SEGMENTATION

Abdominal organ segmentation is one of the most popular segmentation tasks. Most of the existing benchmarks mainly focus on fully supervised segmentation tasks and are built 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 [24]) has achieved very high accuracy. In this section, we evaluate the SOTA method on our plus datasets to show whether the performance can generalize to multi-center datasets.

4.1 Single organ segmentation

Existing 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 the 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 randomly select 80% of cases for training in the original training set and the remaining 20% of cases and three new datasets as testing set, which can allow quantitative comparisons within-dataset and across-dataset.

TABLE 3: 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. The bold and underlined numbers denote the best and worst results, respectively.

Task	Training	Testing	DSC (%)	NSD (%)
Liver	LiTS (104)	LiTS (27)	**97.4±0.63**	83.2±5.89
		KiTS (210)	94.9±7.59	83.2±12.2
		Spleen (41)	96.5±3.31	**86.6±7.54**
		Pancreas (361)	96.4±3.07	85.4±8.46
Kidney	KiTS (168)	KiTS (42)	**97.1±3.81**	**94.0±6.91**
		LiTS (131)	87.5±17.9	75.0±16.5
		Pancreas (361)	82.0±28.9	75.0±27.1
		Spleen (41)	93.7±6.52	82.5±9.97
Spleen	Spleen (33)	Spleen Ts (8)	**97.2±0.81**	**94.6±4.41**
		LiTS (131)	91.0±15.5	79.6±16.4
		KiTS (210)	86.6±23.3	76.7±23.7
		Pancreas (361)	94.6±8.32	86.9±10.4
Pancreas	MSD Pan. (225)	MSD Pan. (56)	86.1±6.59	66.1±15.4
		LiTS (131)	86.6±12.2	75.4±14.2
		KiTS (210)	80.9±10.5	61.5±12.2
		Spleen (41)	**86.6±8.80**	**77.7±11.6**

TABLE 4: Quantitative results of fully supervised multi-organ segmentation in terms of average DSC and NSD. The bold and underlined numbers denote the best and worst results, respectively.

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
	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
	Pancreas Plus (361)	**98.2±1.39**	91.9±5.77	**96.0±5.04**	**92.4±7.06**	97.5±5.88	96.0±7.22	81.1±10.7	61.4±13.3
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
	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
	Pancreas Plus (361)	98.0±2.67	91.3±6.55	94.4±5.61	84.3±9.22	96.8±6.24	94.9±7.96	80.5±11.5	61.5±16.9
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
Spleen Plus (41)	LiTS Plus (131)	96.0±3.35	78.2±7.50	95.2±5.88	85.9±7.19	95.3±9.85	90.6±11.6	88.8±9.61	79.9±11.8
	Pancreas Plus (361)	97.9±2.43	91.2±6.21	95.4±5.45	87.5±6.69	97.7±3.63	96.1±5.75	80.2±12.5	60.7±14.0
	KiTS Plus (210)	96.8±4.80	89.7±9.77	89.7±16.5	85.0±15.0	93.7±13.1	86.4±15.9	83.3±10.7	68.9±12.1

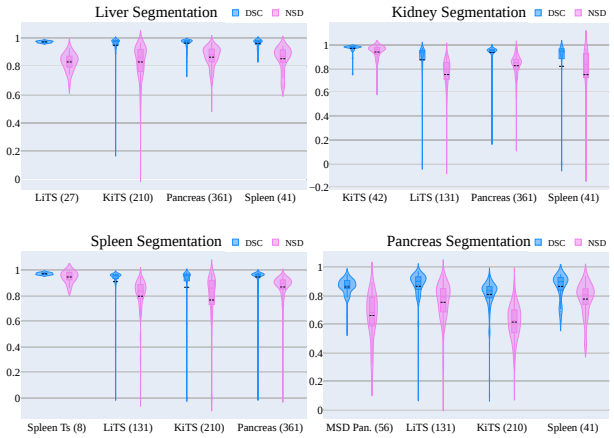


Fig. 6: Violin plots of the segmentation performances (DSC and NSD) of different organs in single organ segmentation tasks.

Table 3 shows the quantitative segmentation results for each organ and Figure 6 shows the corresponding violin plots. It can be found that

- for liver segmentation, the SOTA method achieves high DSC scores ranging from 94.9% to 96.5% on the three new testing datasets, demonstrating its good generalization ability. Compared to the DSC scores on LiTS (27) testing set, the DSC scores drop 2.5% on the KiTS (210). The main reason is that the CT scans in KiTS (210) were acquired on the arterial phase while most of the CT scans in LiTS were acquired on the portal phase. Both Pancreas (361) and Spleen (41) obtain relatively close DSC scores compared with the LiTS (27), but the NSD scores are much better, indicating that the segmentation results in LiTS (27) have more errors near the boundary. This is because most cases in LiTS have liver cancers while most cases in Pancreas (361) and Spleen (41) are normal in liver.
- for kidney segmentation, compared with the high DSC and NSD scores on KiTS (42), the performance drops remarkably on the other three datasets with up to 15% in DSC and 19% in NSD, especially for the LiTS (131) and the Pancreas (361). The main reason is that the CT phases of most cases in the other three datasets are different from the KiTS.
- for spleen segmentation, both DSC and NSD scores also drop on the other three datasets, especially for the KiTS (210) datasets where 10.6% dropping in DSC and 17.9% dropping in NSD is observed, indicating that the SOTA method does not generalize well on different CT phases.
- for pancreas segmentation, the performance also has a significant decline on KiTS (210) because of the differences in CT phases. Remarkably, the LiTS (131) and Spleen (41) obtain similar DSC scores compared to the MSD Pan. (56), but the NSD scores have large improvements with 9.3% and 11% because most cases in the two datasets have a healthy pancreas. The results demonstrate that the pancreas segmentation model generalizes better on pancreas healthy cases than pancreas pathology cases, especially for the boundary-based metric NSD.

In summary, the current SOTA single organ segmentation method can achieve very high performance (especially for the DSC) when the training set and the testing set are from the same distribution, but the high performance would degrade when the testing sets are from new medical centers.

4.2 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 four groups of experiments. In each group, we train the nnU-Net on one dataset with four organ annotations and test the trained model on the other three new datasets. It should be noted that the training set and testing set are from different medical centers in each group.

Table 4 shows quantitative segmentation results for each organ[10]. It can be observed that

- the DSC scores are relatively stable in liver and spleen segmentation results, achieving $90\%+$ in all experiments. However, the NSD scores fluctuate greatly among different testing sets, ranging from 77.4% to 92.1% for liver segmentation and from 86.0% to 97.0% for spleen segmentation.

10. The corresponding violin plots are presented in supplementary (Supplementary Figure 1).

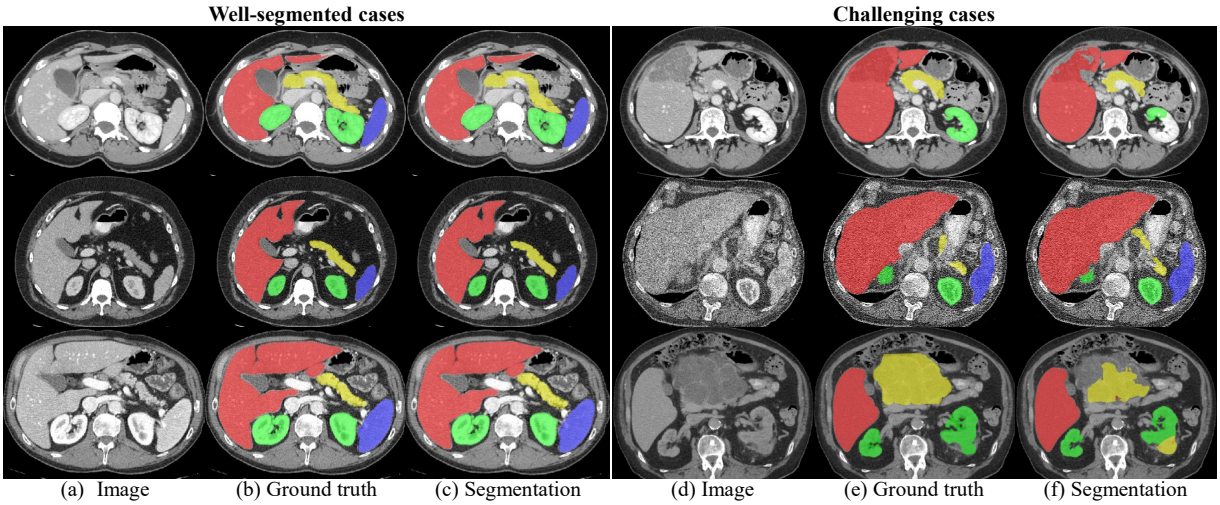


Fig. 7: Well-segmented and challenging examples from testing sets in the large-scale fully supervised multi-organ segmentation study.

- 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 (361) and Spleen Plus (41) datasets, respectively, which has a performance gap of 10%+.
- pancreas segmentation results are lower than the other organs across all experiments, indicating that pancreas segmentation is still a challenging problem.

Figure 7 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 exist no 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 7 (d)-1^{st} row) and the pancreas lesions (Figure 7 (d)-3^{rd} row). In addition, the image quality can be degraded by the noise, e.g., Figure 7 (d)-2^{nd} row.

4.3 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 the testing set is the same as the training set.
- the cases in the testing set are trivial, which means that the cases do not have severe diseases and low image quality.

However, we argue that **abdominal organ segmentation 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 new medical centers with different data distributions from the training set.
- the cases in the 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 2.2, existing abdominal organ segmentation benchmarks cannot reflect these challenging situations. Thus, in this work, we build new segmentation benchmarks that can cover these challenges. Existing benchmarks have received extensive attention in the community and have little rooms for improvements in current testing sets and associated evaluation metric (i.e., DSC) [16], [17]. Therefore, we expect that our new segmentation benchmarks would bring new insights and again attract wide attention.

5 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 planning for organ transplantation.
- evaluating the generalization ability of segmentation methods on cases from new medical centers and CT phases.
- evaluating the generalization ability of segmentation methods on cases with unseen and severe diseases.

In addition to the fully supervised segmentation benchmark, we also set up, to the best of our knowledge, the

first abdominal organ segmentation benchmarks for semi-supervised learning, weakly supervised learning, and continual learning, which are currently active research topics and can alleviate the dependency on annotations. In each benchmark, we select 50 challenging cases and 50 random cases as the testing set, which is friendly to future users to evaluate their methods because it does not cost too much time during inference. More importantly, the final performance is not easy to be biased by the easy cases. We also introduce a new dataset as the common testing set, which can allow apple-to-apple comparisons among the four benchmarks. Moreover, for each benchmark, 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.

5.1 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 4.

5.1.1 Task setting

Motivation of the training set and the testing set choice: a large training set is often expected in fully supervised organ segmentation. Thus, we choose MSD Pan. Plus (281) as the base dataset in the training set because it has the largest number of training cases. On top of MSD Pan. Plus (281), different cases are added to the training set to build two subtasks as shown in Table 5.

- **Subtask 1.** The training set is composed of MSD Pan. Plus (281) and NIH Pan. Plus (80) where all the CT scans are from the portal phase. We use the baseline model in Section 5.1.2 to predict all the remaining cases in LiTS Plus, KiTS Plus, and Spleen Plus. Then, 50 cases with the lowest average DSC and NSD are selected as the testing set. These cases usually have heterogeneous lesions and unclear boundaries, which are very challenging to segment and also very important in clinical practice.
- **Subtask 2.** The added NIH Pan. Plus (80) is replaced by 40 cases from LiTS Plus and 40 cases from KiTS Plus that have similar phases as the testing set. In this way, one can evaluate whether including shared contrast phases across training and testing sets can improve the performance or not.

We use the baseline model in Section 5.1.2 to infer all the remaining cases and select 100 cases as the final testing set, including 50 challenging cases with the lowest average DSC and NSD scores and 50 randomly selected cases. More importantly, the cases in the training set and the testing set have no overlap in each subtask.

5.1.2 Baseline and results

The baseline is built on 3D nnU-Net [24], which is the SOTA method for multi-organ segmentation. Table 5 presents the detailed results for each organ in each subtask. It can be found that the performances of all organs in subtask 1 are lower than the performances in subtask 2 because the cases with shared contrast phases are introduced in subtask 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 [17], [87], our studies on a large and diverse dataset demonstrate that abdominal organ segmentation is still an unsolved problem, especially for the challenging cases and situations.

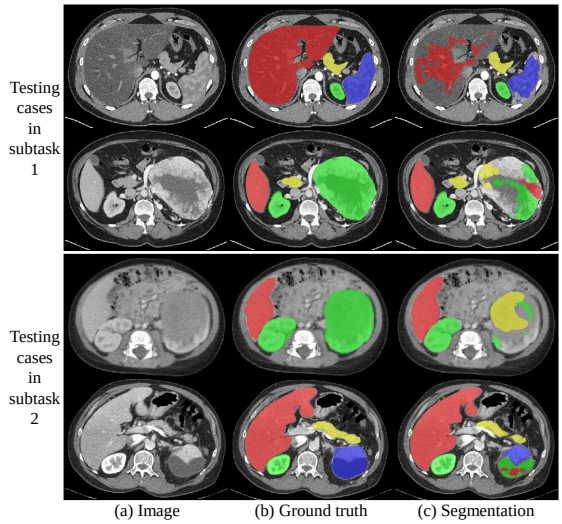


Fig. 8: Challenging examples from testing sets in fully supervised segmentation benchmark.

The violin plots of each organ are presented in Supplementary Figure 2. For the DSC score, though the high DSC scores and low dispersed distributions from the violin plots of the liver segmentation indicate great performance, the results degrade for the other organs. For the NSD score, the obtained scores and the dispersed distributions observed from the violin plots indicate unsatisfying segmentation performance for all four organs. It is worth pointing out that for liver segmentation, the DSC scores are above 95% for both subtasks, indicating great segmentation performance in terms of region overlap between the ground truth and the segmented region. NSD scores are 83% and 85.8% for the two subtasks respectively, demonstrating that the boundary regions contain more segmentation errors, which need further improvements. This phenomenon further proves the necessity of applying NSD for the evaluation of segmentation results.

Figure 8 shows segmentation results of some challenging examples from each subtask. It can be found that the SOTA method does not well generalize to lesion-affected organs. For example, the first row in Figure 8 shows a case with fatty liver in which the liver is darker than the healthy cases. The SOTA method fails to segment the liver completely. The spleen (blue) segmentation result is also poor in this situation. Moreover, the cases in the 2^{nd}, 3^{rd}, and 4^{th} rows have kidney (green), and spleen (blue) tumors, respectively. There exist serious under-segmentation and incorrect segmentation in the segmentation results. These challenging cases are still unsolved problems for abdominal

TABLE 5: 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 (%)
MSD Pan. Plus (281) NIH Pan. Plus (80) Subtask 1: 361 cases	100 cases	95.8±6.04	83.0±12.1	84.1±14.8	73.8±14.1	89.8±15.5	80.6±18.3	65.0±22.7	55.2±17.6
MSD Pan. Plus (281) LiTS Plus (40) KiTS Plus (40) Subtask 2: 361 cases		**97.0±2.93**	**85.8±9.92**	**91.7±11.6**	**84.1±13.1**	**93.6±13.3**	**87.4±15.0**	**78.1±15.8**	**65.0±15.2**

organ segmentation, which are not highlighted in current publicly available benchmarks.

TABLE 6: Task settings of semi-supervised multi-organ segmentation.

Training		Testing	Note
Labelled	Unlabelled		
Spl. Plus (41)	-	100 cases	Lower Bound
Spl. Plus (41)	Pan. Plus (400)		Subtask 1
Spl. Plus (41)	Pan. Plus (400) LiTS Plus (145) KiTS Plus (250) Spl. Plus Ts (5)		Subtask 2
Spl. Plus (41) Pan. Plus (400) LiTS Plus (145) KiTS Plus (250) Spl. Plus Ts (5)	-		Upper Bound

5.2 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 the natural image/video segmentation domain [94], [95]. However, there still exists no related benchmark in the 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.

5.2.1 Task setting

Motivation of the training set and the testing set choice: this semi-supervised task employs MSD Spleen Plus, LiTS Plus, KiTS Plus, MSD Pancreas Plus, and NIH Pancreas Plus as the training and testing datasets. The semi-supervised task is dedicated to alleviating the burden of manual annotations. In this scenario, a small portion of labelled data and a large amount of unlabelled data are available. Therefore, we set the smallest subset, Spleen Plus with 41 cases, as the labelled training set. To show the superiority of semi-supervised methods for leveraging a large amount of unlabelled data, approximately 10-20 times amount of data (400-800 cases) from the remaining subsets are selected as the unlabelled training set. We use the baseline model in Section 5.2.2 to infer all the remaining cases and select 100 cases as the final testing set, including 50 challenging cases with the lowest average DSC and NSD scores and 50 randomly selected cases.

Table 6 presents the semi-supervised segmentation benchmark settings that consist of 2 subtasks. As a contrast, we start with a fully supervised lower-bound task, where a model is trained solely on MSD Spleen Plus containing 41 well-annotated cases. The upper-bound task is also fully supervised that involves the additional 800 labelled cases. Precisely, in upper-bound training set, 41 cases are from MSD Spleen Plus, 400 cases are from MSD and NIH Pancreas Plus, 145 cases are from LiTS Plus, 250 cases are from KiTS Plus, and 5 cases are from MSD Spleen Plus testing set. Based on the lower-bound and upper-bound subtasks, unlabelled cases are gradually involved in the following semi-supervised subtasks. In order to evaluate the effect of the unlabelled data and their quantity on multi-organ segmentation, we carefully design 2 subtasks concerning the source and quantity of unlabelled data. Specifically, subtask 1 utilizes 400 unlabelled cases from MSD and NIH Pancreas Plus, and in addition to the 400 cases, subtask 2 exploits additional 400 unlabelled cases from LiTS plus, KiTS plus, and MSD Spleen Plus testing set. Both subtasks are evaluated on the consistent hold-out testing set for fair comparisons.

5.2.2 Baseline and results

Motivated by the success of the noisy-student learning method in semi-supervised image classification [96] and semi-supervised urban scene segmentation [97] tasks, we develop a teacher-student-based method for semi-supervised abdominal organ segmentation, which includes five main steps:

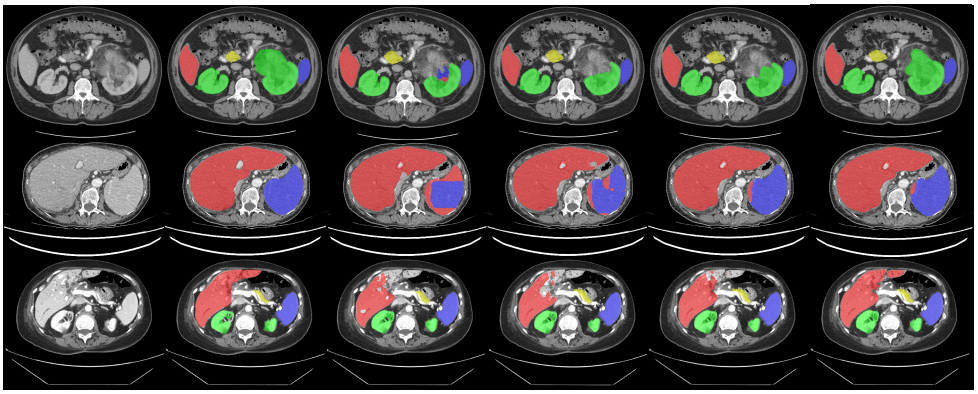
- Step 1. Training a teacher model on the manually labelled data.
- Step 2. Generating pseudo labels of the unlabelled data via the teacher model.
- Step 3. Training a student model on both manual and pseudo labelled data.
- Step 4. Finetuning the student model in step 3 on the manually labelled data.
- 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. The results are presented in Table 7. Due to the different quantity of labelled cases during training, there exists a performance gap between the lower-bound and the upper-bound subtasks. With unlabelled data involved, the performance gradually increased in terms of the average DSC and NSD, indicating that the proposed method can leverage unlabelled cases to improve the multi-organ segmentation performance.

Figure 9 illustrates segmentation results of 3 challenging examples from each subtask. It is observed that our semi-supervised method is able to reduce misclassification by

TABLE 7: Quantitative multi-organ segmentation results in semi-supervised benchmark.

Task	Liver		Kidney		Spleen		Pancreas		Average	
	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
Lower Bound	95.7±5.3	83.0±11	91.3±14	83.8±12	93.6±12	88.2±15	81.5±15	67.7±16	90.5±13	80.7±16
Subtask 1	96.2±4.2	84.0±9.7	91.5±12	83.6±12	94.6±11	90.4±14	82.8±13	69.2±16	91.3±12	81.8±15
Subtask 2	96.2±4.0	83.7±9.5	92.2±12	84.3±11	94.9±11	90.6±13	82.9±13	68.4±15	91.5±12	81.8±15
Upper Bound	**97.4±2.3**	**86.7±8.6**	**95.4±4.0**	**86.9±8.3**	**96.0±9.5**	**92.9±12**	**85.7±8.9**	**72.5±13**	**93.6±8.3**	**84.7±13**



(a) Image (b) Ground Truth (c) Lower Bound (d) Subtask 1 (e) Subtask 2 (f) Upper Bound

Fig. 9: Challenging examples from testing sets in semi-supervised segmentation benchmark.

leveraging unlabelled data. The first and third rows show cases with a large kidney tumor and cholangiectasis inside the liver, respectively. The pathology changes pose an extreme challenge for the kidney segmentation. The second row demonstrates a case that the spleen shares similar appearances with the liver, where it tends to be recognized as the liver when the training data is limited. We can also find that the segmentation error can be gradually corrected by utilizing more unlabelled data. The violin plots of the segmentation results in Supplementary Figure 3 show that a performance increasing trend is observed for the four organs when the quantity of the unlabelled data is increased.

5.3 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 the most commonly used weak annotations for organs segmentation when radiologists manually delineate the organs [88]. In this benchmark, we provide slice-level sparse labels in the training set, where only part ($\leq 30\%$) of the slices are well annotated.

5.3.1 Task settings

Motivation of the training set and the testing set choice: we select the Spleen Plus (41) as the training set because it has the least training cases. This choice is more in line with reality compared with using other datasets (e.g., KiTS Plus (210), LiTS Plus (131)), because the training set has only limited well-annotated cases in many medical centers.

The weakly supervised organ segmentation benchmark contains three subtasks as shown in Table 8, in which 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 [88]. Specifically, we set three different annotation rates 5%, 15%, and 30%, which are similar to the existing work [98] on brain tissue segmentation. We use the baseline model in Section 5.3.2 to infer all the remaining cases and select 100 cases as the final testing set, including 50 challenging cases with the lowest average DSC and NSD scores and 50 randomly selected cases.

TABLE 8: Task settings and quantitative baseline results of weakly supervised abdominal organ segmentation.

Training	Ratio	Testing	DSC (%)	NSD (%)
Spleen Plus (41)	5%	100 cases	78.0 ± 21.8	63.5 ± 20.2
	15%		83.9 ± 17.5	70.4 ± 18.1
	30%		**84.7 ± 16.7**	**70.9 ± 17.6**

5.3.2 Baseline and results

Our baseline method is built on the combination of 2D nnU-Net [24] and fully connected Conditional Random Fields (CRF) [99], which is motivated by the method proposed in [100] where the missing annotation challenge was addressed. The main idea in [100] is to train a pixel-wise classification (segmentation) network with limited labelled images and then segment the unlabelled image to obtain initial segmentation results, followed by a refinement step with fully connected CRF. Fully connected CRF has been widely used in many segmentation tasks (e.g., liver and liver tumor segmentation [101], [102], brain tumor segmentation [103]), which could be an effective way to refine segmentation results. Our new baseline also follows this idea and has the following three main steps:

TABLE 9: Quantitative multi-organ segmentation results in weakly supervised benchmark.

Task	Method	Liver		Kidney		Spleen		Pancreas	
		DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
5% labels	2D U-Net	92.5 ± 6.50	72.7 ± 12.0	**80.3 ± 19.2**	**68.9 ± 15.1**	**82.0 ± 21.7**	69.0 ± 22.5	**57.2 ± 18.9**	**43.3 ± 14.4**
	2D U-Net + CRF	**92.7 ± 6.19**	**72.9 ± 12.1**	78.3 ± 19.7	65.1 ± 15.2	81.8 ± 22.7	**70.3 ± 23.6**	55.2 ± 19.6	42.5 ± 15.2
15% labels	2D U-Net	93.5 ± 6.15	76.5 ± 10.9	**85.0 ± 17.1**	**75.4 ± 15.1**	88.7 ± 15.7	76.6 ± 18.8	**68.5 ± 17.5**	**52.9 ± 14.8**
	2D U-Net + CRF	**93.7 ± 5.89**	**77.2 ± 10.7**	83.4 ± 17.5	71.3 ± 16.0	**89.0 ± 16.2**	**79.0 ± 18.8**	67.2 ± 18.6	52.5 ± 16.2
30% labels	2D U-Net	93.6 ± 6.07	76.6 ± 11.1	**86.0 ± 16.1**	**75.8 ± 14.5**	88.8 ± 15.4	76.1 ± 19.0	**70.5 ± 17.0**	**55.0 ± 14.7**
	2D U-Net + CRF	**93.8 ± 5.76**	**76.9 ± 11.1**	84.3 ± 16.5	72.0 ± 15.3	**89.1 ± 15.9**	**78.4 ± 19.4**	69.3 ± 18.1	54.5 ± 16.1

- Step 1. Training a 2D U-Net [24] with the sparse labels;
- Step 2. Obtaining segmentation probability maps by inferring the testing cases;
- Step 3. Refining the segmentation results with fully connected CRF where the unary potential is the probability map and the pairwise potentials are three Gaussian-kernel potentials defined by the CT attenuation scores [99], [100].

Table 8 presents the average DSC and NSD scores for the four organs, and Table 9 presents the detailed segmentation results for each organ. As expected, the higher annotation ratio the training cases have, the better segmentation performance the baseline method can achieve. With only 15% annotations, the baseline can achieve an average DSC score over 90% for liver segmentation. The results could motivate us to employ deep learning-based strategies to reduce manual annotation efforts and time. In Supplementary Figure 4, we show violin plots of the segmentation results with different annotation ratios. The performance gains from 15% to 30% annotation ratio are fewer than the gains from 5% to 15%, indicating that naively adding annotations cannot always bring linear performance improvements.

In addition, it can be found that using CRF does not bring remarkable performance improvements. A similar phenomenon has also been found by the winner solution [104] in the well-known brain tumor segmentation (BraTS) challenge 2018. Although the results are not promising as expected, they offer new opportunities and challenges for traditional energy-based segmentation methods. Specifically, given initial (inaccurate) CNN segmentation results, how or what kind of energy-based models can consistently improve the segmentation accuracy? All the related results, including trained models, the used CRF code and hyper-parameter settings, the segmentation results and its probability maps will be publicly available for future research along this direction.

5.4 Continual learning benchmark for abdominal organ segmentation

Continual learning has been a newly emerging research topic and attracted significant attention [81]. 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 tasks, e.g., continual learning, incremental learning, life-long learning or online learning. In this paper, we use continual learning to denote such tasks, which is widely used in existing literatures [81]. In CVPR 2020, the first continual learning benchmark, to the best of our knowledge, is set up for image classification[11]. 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 segmentation and develop a baseline solution.

5.4.1 Task setting

Motivation of the training set and the testing set choice: the original single organ datasets, including KiTS (210), Spleen (41), and MSD Pancreas (281), are used as training sets. To evaluate the generalization ability of approaches, we choose MSD Pancreas Ts (139) as the training set with only liver annotation rather than LiTS (131) dataset, because the LiTS (131) is a multi-center dataset that would be better to serve as a testing set. We use the baseline model in Section 5.4.2 to infer all the remaining cases and select 100 cases as the final testing set, including 50 challenging cases with the lowest average DSC and NSD scores and 50 randomly selected cases. As shown in Table 10, 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. In a word, this task requires building a multi-organ segmentation model with only single organ annotated training sets. It also should be noted that one cannot access the previous tasks' dataset when switching to a new task. For example, if a kidney segmentation model has been built with the KiTS (210) dataset, this dataset will be not available when augmenting the model to segment the spleen with Spleen (41) dataset.

5.4.2 Baseline and results

Motivated by the well-known learning without forgetting [105], we develop an embarrassingly simple but effective continual learning method as the baseline, which contains the following four steps:

- Step 1. Individually training a liver segmentation nnU-Net [24] model based on the MSD Pancreas Ts (139) dataset.
- Step 2. Using the trained liver segmentation model to infer KiTS (210) and obtain pseudo liver labels. Thus, each case in the KiTS (210) has both liver and kidney labels. Then, we use the new labels to train a nnU-Net model that can segment both liver and kidney.
- Step 3. Using the trained model in Step 2 to infer Spleen (41) and obtain both liver and kidney pseudo labels. Thus, each case in the Spleen (210) has liver, kidney, and spleen labels. Then, we use the new

11. https://sites.google.com/view/clvision2020/challenge

TABLE 10: 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	100 cases	Liver, kidney, spleen, and pancreas	80.6±10.1	69.8±9.77

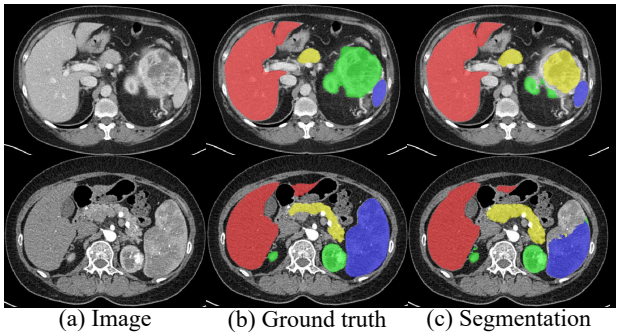
labels to train a nnU-Net model that can segment the three organs.

- Step 4. Using the trained model in Step 3 to infer MSD Pancreas (281) and obtain liver, kidney and spleen pseudo labels. Thus, each case in the MSD Pancreas (281) has liver, kidney, spleen, and pancreas labels. Finally, we can obtain the final multi-organ segmentation model by training a nnU-Net with the new labels.

TABLE 11: Quantitative multi-organ segmentation results of continual learning.

Organ	DSC (%)	NSD (%)
Liver	94.7±7.99	81.7±14.0
Kidney	79.4±18.9	73.6±16.5
Spleen	83.8±23.2	72.9±24.2
Pancreas	64.7±21.6	51.1±16.3

Table 10 presents the average DSC and NSD scores for the four organs, and Table 11 presents the detailed results for each organ. Overall, the performance of learning with single organ datasets is lower than learning with the full annotations as presented in the fully supervised segmentation results (Table 4, Table 5), indicating that the model still tends to forget part of the previous tasks when switching to a new task. The violin plots of the segmentation performance for each organ are presented in Supplementary Figure 5. Liver segmentation obtains the best DSC and NSD scores with compact distributions and fewer outliers while pancreas segmentation obtains lower performance. The low scores and dispersed distributions of NSD reveal relatively high boundary errors because of the effects of various pathological changes as shown in Figure 10.



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

Fig. 10: Challenging examples from testing sets in continual learning multi-organ segmentation benchmark.

5.5 Evaluation and comparison on the common testing set

The above testing sets are different in the four benchmarks. For an apple-to-apple comparison between the benchmarks, we introduce a common testing set (NJU dataset as described in Section 3.1) with 50 abdomen CT cases.

Table 12 presents the quantitative results of the common testing set of the four benchmarks. It is observed that the fully supervised method achieves the best average DSC and NSD scores for kidney, spleen and pancreas segmentation, because it uses many labelled cases. However, due to the burden of annotation, it is usually difficult to obtain the desired amount of annotations in clinical practice. Therefore, the problem lies in that what is the desired annotation-efficient method. The semi-supervised method in subtask 2 with only 41 labelled cases and 800 unlabelled cases achieves the best performance for the liver and the overall performance is very close to the best fully supervised method (361 labelled cases), indicating that using large and diverse unlabelled cases can significantly improve the performance. The weakly supervised methods achieve the lowest performance, but it requires the least annotation burden.

6 Conclusion

In this work, we have introduced AbdomenCT-1K, the largest abdominal CT organ segmentation dataset, which includes multi-center, multi-phase, multi-vendor, and multi-disease cases. Although the SOTA method has achieved unprecedented performance in several existing benchmarks, such as liver, kidney, and spleen segmentation, our large-scale studies reveal that some problems remain unsolved as shown in Section 4. In particular, the SOTA method can achieve superior segmentation results when the evaluation metric is DSC, the testing set has a similar data distribution as the training set, and no hard cases with unseen diseases in the testing set. However, the SOTA method cannot generalize the great performance on unseen datasets with many challenging cases, such as the cases with new CT phases, severe diseases, acquired from distinct scanners or clinical centers.

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 (e.g., LiTS [16], MSD [20], and KiTS [17]), our new benchmarks have three main characteristics:

- 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

TABLE 12: Quantitative results on the common testing set of the four benchmarks.

Task		Liver		Kidney		Spleen		Pancreas		Average	
		DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
Fully Supervised	Subtask 1	95.9±5.4	87.5±8.7	94.8±6.5	89.2±12	86.3±18	78.2±22	76.3±24	65.1±22	88.3±17	80.0±20
	Subtask 2	97.5±2.5	89.3±5.2	**97.4±3.8**	**95.7±6.6**	**97.3±5.4**	**94.5±10**	**82.5±19**	**71.8±18**	**93.6±12**	**87.8±15**
Semi-Supervised	Subtask 1	96.9±1.4	83.2±7.1	96.0±4.4	90.0±7.6	93.3±11	86.3±18	72.5±20	57.9±19	89.7±15	79.4±19
	Subtask 2	**97.9±1.0**	**91.2±4.0**	97.1±4.3	93.4±6.8	97.2±3.9	94.1±9.7	82.3±12	70.7±13	**93.6±9.4**	87.4±13
Weakly Supervised	Subtask 1	84.8±9.8	55.0±10	73.7±24	56.1±21	63.8±34	55.5±28	16.8±19	14.8±16	60±35	45.6±27
	Subtask 2	84.8±9.7	55.3±11	81.6±17	62.4±19	68.7±31	58.6±27	29.8±21	20.4±17	66.2±30	49.2±25
	Subtask 3	84.8±9.4	55.4±11	83.1±11	62.7±17	68.8±29	57.5±26	30.6±22	21.4±18	66.8±29	49.2±25
Continual Learning		93.6±9.8	79.6±13	90.3±13	81.7±16	80.8±24	67.1±22	77.0±20	59.1±19	85.4±19	71.9±20

in the preoperative planning of many abdominal organ surgeries, such as tumor resections and organ transplantation.

The main limitation is that we only focus on the segmentation of four large abdominal organs. However, there exist far more difficult organs [64], [106] and lesions where the annotations are not available in our dataset. To address this limitation, we annotate 50 cases with 8 extra organs, including esophagus, gallbladder, stomach, aorta, celiac trunk, inferior vena cava, right adrenal gland and left adrenal gland. For the lesions, the detailed pathology information is not available in the original dataset. It is challenging to make a definite and accurate diagnosis with only CT scans because identifying the (malignant) tumor usually requires pathological examinations. As an alternative, we include the original single-organ tumor masks [16], [17], [20] and provide pseudo tumor labels of 663 cases by annotating all the other possible tumors, which can be used for noisy label learning.

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 baseline methods help push abdominal organ segmentation towards the real clinical practice.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *Medical Image Analysis*, vol. 70, p. 101979, 2021.

[14] "DAVIS: Densely Annotated VIdeo Segmentation," https://davischallenge.org/, 2020, [Online; Accessed: Aug. 2020].

[15] 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.

[16] 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.

[17] 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*, vol. 67, p. 101821, 2021.

[18] 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.

[19] 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.

[20] 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.

[21] H. R. Roth, A. Farag, E. B. Turkbey, L. Lu, J. Liu, and R. M. Summers, "Data from pancreas-ct," The Cancer Imaging Archive, 2016.

[22] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 556–564.

[23] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[24] F. Isensee, P. F. Jäeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[25] 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.

[26] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] J. Yao, J. Cai, D. Yang, D. Xu, and J. Huang, "Integrating 3d geometry of organ for improving medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 318–326.

[41] K. George, A. P. Harrison, D. Jin, Z. Xu, and D. J. Mollura, "Pathological pulmonary lobe segmentation from ct images using progressive holistically nested neural networks and random walker," in *Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 195–203.

[42] A. P. Harrison, Z. Xu, K. George, L. Lu, R. M. Summers, and D. J. Mollura, "Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2017, pp. 621–629.

[43] D. Jin, D. Guo, T.-Y. Ho, A. P. Harrison, J. Xiao, C.-k. Tseng, and L. Lu, "Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy," *Medical Image Analysis*, vol. 68, p. 101909, 2021.

[44] 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.

[45] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical Image Analysis*, vol. 45, pp. 94–107, 2018.

[46] H. R. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2016, pp. 451–459.

[47] 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.

[48] 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.

[49] Z. Zhu, C. Liu, D. Yang, A. Yuille, and D. Xu, "V-nas: Neural architecture search for volumetric medical image segmentation," in *2019 International Conference on 3D Vision*, 2019, pp. 240–248.

[50] D. Guo, D. Jin, Z. Zhu, T.-Y. Ho, A. P. Harrison, C.-H. Chao, J. Xiao, and L. Lu, "Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4223–4232.

[51] Y. He, D. Yang, H. Roth, C. Zhao, and D. Xu, "Dints: Differentiable neural network topology search for 3d medical image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[52] 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.

[53] 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.

[54] 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*, 2016, pp. 111–120.

[55] 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.

[56] 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.

[57] 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*, 2018, pp. 417–425.

[58] 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.

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

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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, 2020, p. 113131I.

[66] A. Raju, C.-T. Cheng, Y. Huo, J. Cai, J. Huang, J. Xiao, L. Lu, C. Liao, and A. P. Harrison, "Co-heterogeneous and adaptive segmentation from multi-source and multi-phase ct imaging data: a study on pathological liver and lesion segmentation," in *European Conference on Computer Vision*, 2020, pp. 448–465.

[67] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Medical Image Analysis*, vol. 65, p. 101766, 2020.

[68] 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*, 2017, pp. 79–87.

[69] 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*, 2019, pp. 1409–1415.

[70] 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.

[71] 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*, 2016, pp. 549–565.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] X. Wang, S. Liu, H. Ma, and M.-H. Yang, "Weakly-supervised semantic segmentation by iterative affinity learning," *International Journal of Computer Vision*, vol. 128, pp. 1736–1749, 2020.

[78] 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.

[79] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgeting in gradientbased neural networks," in *In Proceedings of International Conference on Learning Representations*, 2014.

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

[81] 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.

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

[83] 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*, 2017, pp. 3207–3214.

[84] 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.

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

[86] O. Jimenez-del Toro, H. Müller, M. Krenn, K. Gruenberg, A. A. Taha, M. Winterstein, I. Eggel, A. Foncubierta-Rodríguez, O. Goksel, A. Jakab *et al.*, "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2459–2475, 2016.

[87] 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.

[88] 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.

[89] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin, "Ct-org, a new dataset for multiple organ segmentation in computed tomography," *Scientific Data*, vol. 7, no. 1, pp. 1–9, 2020.

[90] 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.

[91] J. Ma, "Cutting-edge 3d medical image segmentation methods in 2020: Are happy families all alike?" *arXiv preprint arXiv:2101.00232*, 2021.

[92] 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.

[93] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 109–117.
[100] M. Gao, Z. Xu, L. Lu, A. Wu, I. Nogues, R. M. Summers, and D. J. Mollura, "Segmentation label propagation using deep convolutional neural networks and dense conditional random field," in *2016 IEEE 13th International Symposium on Biomedical Imaging*, 2016, pp. 1265–1268.
[101] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi *et al.*, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 415–423.
[102] Y. Zhang, Z. He, C. Zhong, Y. Zhang, and Z. Shi, "Fully convolutional neural network with post-processing methods for automatic liver segmentation from ct," in *2017 Chinese Automation Congress*, 2017, pp. 3864–3869.
[103] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical Image Analysis*, vol. 43, pp. 98–111, 2018.
[104] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*, 2018, pp. 311–320.
[105] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.
[106] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 514–525, 2019.