# Learning Incrementally to Segment Multiple Organs in a CT Image

Pengbo Liu[1,2], Xia Wang[3], Mengsi Fan[3], Hongli Pan[3], Minmin Yin[3], Xiaohong Zhu[3], Dandan Du[3], Xiaoying Zhao[3], Li Xiao[2], Lian Ding[4], Xingwang Wu[3], and S. Kevin Zhou[1,2](✉)

[1] Center for Medical Imaging, Robotics, Analytic Computing and Learning (MIRACLE), School of Biomedical Engineering and Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China
skevinzhou@ustc.edu.cn

[2] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China

[3] The First Affiliated Hospital of Anhui Medical University, Anhui, China

[4] Huawei Cloud Computing Technology Co. Ltd., Dongguan, China

**Abstract.** There exists a large number of datasets for organ segmentation, which are partially annotated and sequentially constructed. A typical dataset is constructed at a certain time by curating medical images and annotating the organs of interest. In other words, new datasets with annotations of new organ categories are built over time. To unleash the potential behind these partially labeled, sequentially-constructed datasets, we propose to incrementally learn a multi-organ segmentation model. In each incremental learning (IL) stage, we lose the access to previous data and annotations, whose knowledge is assumingly captured by the current model, and gain the access to a new dataset with annotations of new organ categories, from which we learn to update the organ segmentation model to include the new organs. While IL is notorious for its 'catastrophic forgetting' weakness in the context of natural image analysis, we experimentally discover that such a weakness mostly disappears for CT multi-organ segmentation. To further stabilize the model performance across the IL stages, we introduce a *light memory module* and some loss functions to restrain the representation of different categories in feature space, aggregating feature representation of the same class and separating feature representation of different classes. Extensive experiments on five open-sourced datasets are conducted to illustrate the effectiveness of our method.

**Keywords:** Incremental learning · Partially labeled datasets · Multi-organ segmentation

## 1 Introduction

While most natural image datasets [3,10] are completely labeled for common categories, fully annotated medical image datasets are scarce, especially for

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13434, pp. 714–724, 2022.
https://doi.org/10.1007/978-3-031-16440-8_68

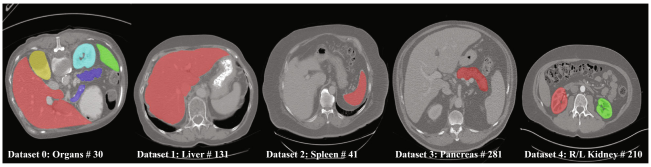

**Fig. 1.** Number of cases in different partially labeled datasets for different tasks.

a multi-organ segmentation (MOS) task [29] that requires pixel-wise annotations, as constructing such a dataset requires professional knowledge of different anatomical structures [28,29]. Fortunately, there exist many partially labeled datasets [1,5,24] for organ segmentation. Another dimension associated with these datasets is that they are constructed sequentially at different sites. Our goal is to train **a single multi-organ segmentation model from partially labelled, sequentially constructed datasets**. To achieve such a goal, we have to address two issues. (i) The first issue arising from *partial labeling* is knowledge conflict, that is, labels in different datasets have conflicts, *e.g.*, the liver is marked as foreground in Dataset 1 but as background in Datasets 2–4, as shown in Fig. 1. (ii) The second issue arising from *sequential construction* is data availability, that is, the datasets are not simultaneously available for learning. What could be even worse is that, due to security concern, these datasets are not allowed to be transferred across the border of the curating institutes; only the model parameters are sharable.

There has been some emerging research [4,21,27,30] that successfully handles knowledge conflict and trains a single model from pooled datasets for improved performance in multi-organ segmentation, proving that the unlabeled data in partially labeled datasets is also helpful for learning. However, these approaches conduct model learning in a batch model based and hence unable to be applied to deal with sequential construction. To deal with both issues, we hereby propose a novel multi-organ segmentation approach based on *the principle of incremental learning (IL)*, which is a staged learning method that has an access to the data available at current learning stage, while losing the access to the data available in previous stages.

Our main contributions are summarized as below:

- We make the first attempt in the literature to merge partially labeled datasets in medical image scenario using IL method, addressing the issues of knowledge conflict and data availability, and possibly security concern.
- To combat the 'catastrophic forgetting' problem that commonly plagues IL, we introduce a light memory module [7] to store the prototypical representation of different organ categories and corresponding loss functions to make different organs more distinguishable in feature space.
- Our extensive experiments on five open-source organ datasets achieve comparable performance to state-of-the-art (SOTA) batch methods which can access all datasets in training phase, unleashing the great potential of IL in multiple organ segmentation.

## 2 Related Work

**MOS with Partially Labelled Datasets.** Zhou *et al.* [30] learn a segmentation model in the case of partial labeling by adding a prior-aware loss in the learning objective to match the distribution between the unlabeled and labeled datasets. In [4], first multi-scale features at various depths are hierarchically incorporated for image segmentation and then a unified segmentation strategy is developed to train three separate datasets together, and finally multi-organ segmentation is achieved by learning from the union of partially labeled and fully labeled datasets. Zhang *et al.* [27] propose a dynamic on-demand network (DoDNet) that learns to segment multiple organs and tumors on partially labeled datasets, which embedded dynamically generated filter by a task encoding module into an encoder-decoder architecture. Shi *et al.* [21] encode knowledge from different organs into a single multi-class segmentation model by introducing two simple but effective loss functions, *Marginal* loss and *Exclusion* loss.

**Incremental Learning.** IL has been studied for object recognition [8,9,11,19] and detection [15,22,23], also segmentation [2,17,18,25]. The main challenge in IL is the so-called 'catastrophic forgetting' [16]: how to keep the performance on old classes while learning new ones? Methods based on parameter isolation [20,26] and data replay [13,19] are all with limited scalability or privacy issues. Regularization based method is the most ideal direction in IL community. In natural image segmentation, Cermelli et al. [2] solved knowledge conflicts existing in other IL methods [9,17] by remodeling old and new categories into background in loss functions, achieving a performance improvement. In 2D medical image segmentation, Ozdemir and Goksel [18] made some attempts using the IL methods used in natural images directly, with only two categories, and it mainly focuses on verifying the possibility of transferring the knowledge learned in the first category with more images to a second category with less images. In this paper, we apply IL to multiple organ segmentation for the first time.

## 3 Method

### 3.1 IL for MOS

**Framework of IL.** The overview of the $t^{th}$ stage of IL in our method is shown in Fig. 2. Given a pair of 3D input image and ground truth, $\{x^t, y^t\} \in \{\mathcal{X}^t, \mathcal{C}^t\}$, we firstly process $x^t$ by the model in current stage, $f_{\theta_t}(\cdot)$ with trainable parameters $\theta_t$, getting the output $q^t = f_{\theta_t}(x^t)$. And we assume that each image $x^t$ is composed by a set of voxels $x_i^t$ with constant cardinality $|\mathcal{I}| = N$. The whole label space $\mathcal{Y}^t$ cross all $t$ stages is expanded from $\mathcal{Y}^{t-1}$ with new classes added in current stage ($\mathcal{C}^t$), $\mathcal{Y}^t = \mathcal{Y}^{t-1} \cup \mathcal{C}^t = \mathcal{C}^1 \cup ... \cup \mathcal{C}^t$. Note that the annotations of the old categories $\mathcal{Y}^{t-1}$ will be inaccessible in the new stage under ideal IL settings. For preserving the knowledge of old categories in regularization based method, we process $x^t$ by the saved old model $f_{\theta_{t-1}}(\cdot)$ with frozen parameters $\theta_{t-1}$ and get $q^{t-1} = f_{\theta_{t-1}}(x^t)$ as the pseudo label. Knowledge distillation loss,

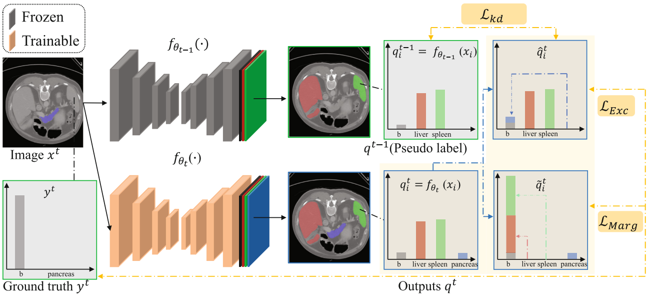

**Fig. 2.** Overview of the $t^{th}$ stage of IL in multi-organ segmentation.

$\mathcal{L}_{kd}$, is introduced in IL setting to keep old knowledge learned from previous stages. Trainable $\theta_t$ in the $t^{th}$ stage is expanded from $\theta_{t-1}$ with $\Theta_t$ to segment new categories, $\theta_t = \theta_{t-1} \cup \Theta_t$.

**Avoiding Knowledge Conflict in IL.** The structures of old classes in $\mathcal{X}^t$, are marked as background in $\mathcal{C}^t$. And the new structures also do not exist in $\mathcal{Y}^{t-1}$, that is new structures are marked as background in pseudo label. If we directly use $q^t$ to compute segmentation loss for new classes, and knowledge distillation loss for old classes, these conflicts between prediction and ground truth break the whole training process. So referring to marginal loss in MargExc [21], we modify the prediction $q^t$ to $\hat{q}^t$ and $\tilde{q}^t$, as shown in Fig. 2 and Eqs. (1) and (2):

$$\hat{q}_{i,j}^t = \begin{cases} \exp(q_{i,b}^t + \sum_{c\in\mathcal{C}^t} q_{i,c}^t)/\sum_{c\in\mathcal{Y}^t\cup b}\exp(q_{i,c}^t) & \mathit{if}\ j = b \\ \exp(q_{i,j}^t)/\sum_{c\in\mathcal{Y}^t\cup b}\exp(q_{i,c}^t) & \mathit{if}\ j \in \mathcal{Y}^{t-1} \end{cases} \tag{1}$$

$$\tilde{q}_{i,j}^t = \begin{cases} \exp(q_{i,b}^t + \sum_{c\in\mathcal{Y}^{t-1}} q_{i,c}^t)/\sum_{c\in\mathcal{Y}^t\cup b}\exp(q_{i,c}^t) & \mathit{if}\ j = b \\ 0 & \mathit{if}\ j \in \mathcal{Y}^{t-1} \\ \exp(q_{i,j}^t)/\sum_{c\in\mathcal{Y}^t\cup b}\exp(q_{i,c}^t) & \mathit{if}\ j \in \mathcal{C}^t \end{cases} \tag{2}$$

where $b$ means background. Then the probability of classes not marked in ground truth or pseudo label will not be broken during training.

### 3.2 Memory Module

As shown in Fig. 3, representation $\mathcal{R}$ is feature maps out of decoder, with shape of C×D×H×W, where C means the number of channels in $\mathcal{R}$. To further mitigate ‘knowledge forgetting’ in IL setting, we introduce a light memory module $\mathcal{M}$ of size $|\mathcal{Y}^t| \times$C in feature space between decoder and segmentation head, $\mathfrak{H}$, to remember the representation of each class. The size of $\mathcal{M}$ is updated by more $|\mathcal{C}^t| \times$C on $\left|\mathcal{Y}^{t-1}\right| \times$C after the $t^{th}$ stage. Then based on $\mathcal{M}$ we can add some constraints in feature space to improve the IL learning progress.

During training of each stage, with the position supplied by ground truth, we can acquire the voxel representation of corresponding new organs in feature

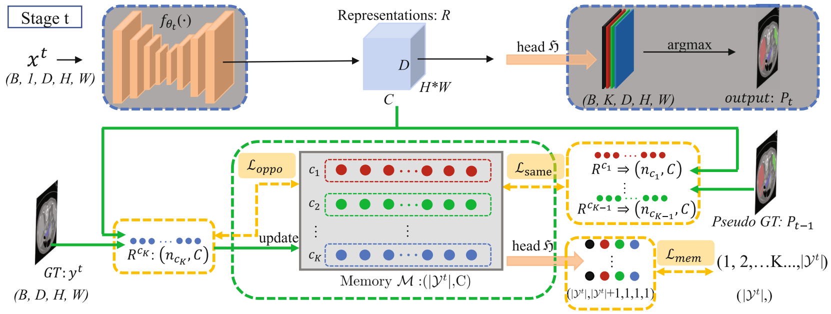

**Fig. 3.** Diagram of the memory module $\mathcal{M}$ in feature space between decoder and segmentation head. Based on label, we can take $n_{c_K}$ voxels' representation from $\mathcal{R}$, $\mathcal{R}^{c_K}$:($n_{c_K}$, C), to update $\mathcal{M}$ or calculate loss function.

map $\mathcal{R}$. Then new class $c$ in $\mathcal{M}$ can be updated via moving average after each iteration:

$$\mathcal{M}_k^c = (1-m_k)\cdot\mathcal{M}_{k-1}^c + m_k\cdot\mathcal{R}_k^c, \quad m_k = \frac{9m_0}{10}\cdot(1-\frac{k}{K})^p + \frac{m_0}{10}, \tag{3}$$

where $m$ is the momentum, $k$ denotes the current number of iterations, and $K$ is the total number of iterations of training. $p$ and $m_0$ are set as 0.9 empirically. After each stage of training ends, the mean representation of new organ of category $c$ in that stage is saved into the memory $\mathcal{M}$ as $\mathcal{M}^c$.

When we have $\mathcal{M}$ to save the mean representation of each class, we can introduce more regularization to constrain the learning of feature space. In this paper, we introduce $l_{mem}$, $l_{same}$ and $l_{oppo}$:

$$\mathcal{L}_{mem} = \mathcal{L}_{ce}(\mathfrak{H}(reshape(\mathcal{M})), range(1, |\mathcal{Y}^t|+1)) \tag{4}$$

$$\mathcal{L}_{same} = \sum_{c_0\in\mathcal{Y}^{t-1}} \mathcal{L}_{cos}(\mathcal{M}^{c_o}, \mathcal{R}^{c_o}, 1) \tag{5}$$

$$\mathcal{L}_{oppo} = \sum_{c_n\in\mathcal{C}^t} (\mathcal{L}_{cos}(\mathcal{R}^b, \mathcal{R}^{c_n}, -1) + \sum_{c_0\in\mathcal{Y}^{t-1}} \mathcal{L}_{cos}(\mathcal{M}^{c_o}, \mathcal{R}^{c_n}, -1)) \tag{6}$$

In Eq. (4), *reshape* is used to change $\mathcal{M}$ to the size of $|\mathcal{Y}^t| \times C \times 1 \times 1 \times 1$, which can be regarded as $|\mathcal{Y}^t|$ voxels belong to $|\mathcal{Y}^t|$ classes. $range(1, |\mathcal{Y}^t|+1)$ can be seen as corresponding ground truth. Through the shared segmentation head $\mathfrak{H}$, features of classes in current stage are going to center around the mean representation in $\mathcal{M}$. Through $l_{mem}$, we constrain the learned feature of different classes in different stages more stable. The mean representation of old classes are treated as a kind of replay without privacy concerns. In Eqs. (5) and (6), $c_o$ and $c_n$ refer to old and new classes, respectively. Using Cosine Embedding Loss, $\mathcal{L}_{cos}$, we can explicitly restrain the feature of old class close to $\mathcal{M}^{c_o}$, and the feature of new class away from all $\mathcal{M}^{c_o}$.

**Table 1.** A summary of five benchmark datasets used in our experiments. [T] means there are tumor labels in original dataset and we merge them into corresponding organs.

| Phase | Datasets | Modality | # of labeled volumes | Annotated organs | Mean spacing (z, y, x) | Source |
|---|---|---|---|---|---|---|
| *Training & Val* | Dataset0 (F) | CT | 30 | Five organs | (3.0, 0.76, 0.76) | Abdomen in [1] |
| | Dataset1 ($P_1$) | CT | 131 | Liver [T] | (1.0, 0.77, 0.77) | Task03 in [24] |
| | Dataset2 ($P_2$) | CT | 41 | Spleen | (1.6, 0.79, 0.79) | Task09 in [24] |
| | Dataset3 ($P_3$) | CT | 281 | Pancreas [T] | (2.5, 0.80, 0.80) | Task07 in [24] |
| | Dataset4 ($P_4$) | CT | 210 | L&R Kidneys [T] | (0.8, 0.78, 0.78) | KiTS [5] |
| | All | CT | 693 | Five organs | (1.7, 0.79, 0.79) | - |
| *Testing* | CLINIC | CT | 107 | Five organs | (1.2, 0.74, 0.74) | Private |
| | Amos | CT | 200 | Five organs | (5.0, 0.74, 0.74) | Temporarily private |
| | Pan | CT | 56 | Five organs | (2.6, 0.82, 0.82) | FLARE 21 [14] |

## 4 Experiments

### 4.1 Setup

**Datasets and preprocessing.** To compare with our base method, MargExc [21], we choose the same five organs and datasets in our experiments, including liver, spleen, pancreas, right kidney and left kidney. In addition, we find three more independent datasets for testing to give a comprehensive evaluation. The details of these datasets are shown in Table 1.

We preprocess all datasets to a unified spacing (2.41, 1.63, 1.63) and normalize them with mean and std of 90.9 and 65.5 respectively. We respectively split five training datasets into 5 folds and randomly select one fold as validation set. For our main IL setting, five organs are learned in four stages: liver (F+$P_1$)→ spleen (F+$P_2$)→ pancreas (F+$P_3$)→ R/L kidney (F+$P_4$). The annotations of different organs in dataset F are used separately in our IL setting.

**Implementation Details.** We implement our experiments based on 3D lowres version of nnU-Net[1] [6] and also refer to MONAI[2] during our algorithm development. The patch-size and batch-size are set as (80, 160, 128) and 2, respectively, in our experiments. We train the network with the same optimizer and learning rate policy as nnU-Net for 350 epochs. The initial learning rate of the first stage and followed stages are set to 3e-4 and 15e-5.

**Baseline Methods.** Intuitively, we train a 5-class segmentation model $\phi_F$ on dataset F directly. And to use more partially labeled datasets, we train 4 models separately for different organs, too, i.e., $\phi_{F+P_*}$. To simulate different organs are collected sequentially, simple fine-tuning (FT) and some SOTA IL methods (LwF [9], ILT [17] and MiB [2]) are also implemented. In the end, to evaluate our performance in actual usage scenarios, we compare our method to the upper bound results from MargExc [21]. Since we also use the marginal loss in IL, we call our method MargExcIL.

[1] github.com/mic-dkfz/nnunet.

[2] https://monai.io/.

**Table 2.** In the last stage($4^{th}$), the DC and HD95 of the segmentation results of different methods. MargExc [21] is the upper bound method training all datasets in the meantime. '-' means no result.

| Training form | Methods | Organs | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DC/HD95 ($F + P_i$) | | | | | | | | | | | |
| | | Liver | | Spleen | | Pancreas | | R Kidney | | L Kidney | | Mean | |
| | | DC | HD | DC | HD | DC | HD | DC | HD | DC | HD | DC | HD |
| *Trained separated* | $\phi_F$ (Five organs) | .953 | 10.28 | .953 | 1.93 | .721 | 8.25 | .895 | 5.82 | .839 | 13.41 | .872 | 7.94 |
| | $\phi_{F+P_1}$ (Liver) | .967 | 5.89 | - | - | - | - | - | - | - | - | .936 | 8.25 |
| | $\phi_{F+P_2}$ (Spleen) | - | - | .954 | 20.20 | - | - | - | - | - | - | | |
| | $\phi_{F+P_3}$ (Pancreas) | - | - | - | - | .842 | 5.13 | - | - | - | - | | |
| | $\phi_{F+P_4}$ (Kidneys) | - | - | - | - | - | - | .968 | 5.18 | .950 | 4.86 | | |
| *One model* | FT | .000 | - | .000 | - | .000 | - | .970 | 6.502 | .963 | 2.018 | .387 | - |
| | LwF [9] | .001 | 190.33 | .906 | 2.22 | .792 | 5.91 | .966 | 7.70 | .948 | 7.22 | .723 | 42.68 |
| | ILT [17] | .000 | 170.77 | .914 | 2.05 | .772 | 8.40 | .969 | 1.41 | .948 | 4.06 | .721 | 37.34 |
| | MiB [2] | .966 | 6.76 | .961 | 1.26 | .817 | 6.56 | .966 | 3.77 | .946 | 7.22 | .931 | 5.11 |
| *Ours* | MargExcIL | .965 | 7.98 | .962 | 1.30 | .835 | 5.51 | .968 | 1.40 | .959 | 2.37 | .938 | 3.71 |
| Upper bound | MargExc [21] | .962 | 7.01 | .965 | 1.15 | .848 | 4.83 | .969 | 1.39 | .965 | 3.96 | **.942** | **3.67** |

**Performance Metrics.** We use Dice coefficient (DC) and $95^{th}$ percentile Hausdorff distance (HD95) to evaluate results.

### 4.2 Results and Discussions

**Comparison with Baseline Methods.** In IL setting, performance of batch learning of all categories is seen as the upper bound for comparison. Because joint learning can access all knowledge at the meantime, it is possible to fit the distribution of the whole dataset. We regard MargExc [21] as the counterpart batch method in MOS, which obtains the DC of 0.942 and HD95 of 3.67 when training all five training datasets together, as in Table 2.

When we do not aggregate these partially labeled data together, there are some limitations in performance. The 5-class segmentation model $\phi_F$ only trained on small scale 'fully' annotated dataset F, can not generalize well to all validation datasets due to the scale of the dataset F. The metrics of DC and HD95 are all much worse than upper bound. When we train four models, $\phi_{F+P_*}$, one model per organ segmentation task trained on corresponding datasets (F+$P_*$), then all datasets can be used. We can get much better performance than $\phi_F$ on DC metric, but also bad HD95 metric. Higher HD95 means more false positive predictions out of our trained models. Furthermore, training separately is also poor in scalability and efficiency when the categories grow in the future.

When we aggregate these partially labeled datasets together sequentially, the most intuitive method FT is the worst. It has no preservation of the old knowledge because there is no restraint for it. LwF [9] and ILT [17] perform better than FT, but 'knowledge conflict' limits the performance of LwF and ILT when the stage of IL is more than 3, *i.e.*, the liver knowledge in the $1^{st}$ stage can not be kept in the $4^{th}$ stage. We check the output of the models trained via LwF

**Table 3.** The DC and HD95 of the segmentation results. The best and second result is shown in **bold** and red. S$*$ means $*^{th}$ stage in IL setting. '-' means **No Access** to the classes in that stage.

| Setting | Organs | DC/HD95 | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Model$_{S*}$ on Validation Sets | | | | | | | | Model$_{S3}$ on Testing Datasets | | | | | |
| | | S0 | | S1 | | S2 | | S3 | | CLINIC | | AMOS | | Pan | |
| | | DC | HD | DC | HD | DC | HD | DC | HD | DC | HD | DC | HD | DC | HD |
| *MargExc−IL*$_{swin}$ *(woMem)* | Liver | .965 | 5.51 | .959 | 20.10 | .958 | 20.86 | .957 | 7.24 | .971 | 3.33 | .937 | 10.68 | .977 | 2.00 |
| | Spleen | - | - | .958 | 2.11 | .960 | 2.12 | .962 | 1.19 | .953 | 5.88 | .865 | 6.93 | .965 | 1.66 |
| | Pancreas | - | - | - | - | .827 | 6.02 | .809 | 6.85 | .853 | 5.80 | .610 | 21.91 | .809 | 6.85 |
| | R Kidney | - | - | - | - | - | - | .966 | 1.46 | .945 | 6.80 | .837 | 7.10 | .942 | 3.88 |
| | L Kidney | - | - | - | - | - | - | .959 | 2.29 | .941 | 6.83 | .872 | 7.64 | .948 | 4.41 |
| | mean | .965 | 5.51 | .959 | 11.11 | .915 | 9.67 | .931 | 3.81 | .933 | 5.73 | .824 | 10.85 | .928 | 3.76 |
| *MargExc−IL*$_{swin}$ | Liver | .965 | 3.60 | .962 | 7.33 | .959 | 8.08 | .958 | 9.04 | .970 | 3.46 | .950 | 4.08 | .974 | 2.42 |
| | Spleen | - | - | .961 | 1.24 | .964 | 1.18 | .963 | 1.15 | .953 | 3.50 | .882 | 6.00 | .966 | 1.58 |
| | Pancreas | - | - | - | - | .826 | 5.21 | .816 | 5.65 | .847 | 6.24 | .640 | 31.31 | .817 | 5.71 |
| | R Kidney | - | - | - | - | - | - | .964 | 3.49 | .942 | 7.31 | .883 | 6.43 | .941 | 4.63 |
| | L Kidney | - | - | - | - | - | - | .953 | 2.59 | .933 | 8.94 | .867 | 8.96 | .951 | 3.60 |
| | mean | .965 | 3.60 | .962 | 4.28 | .916 | 4.821 | .931 | 4.38 | .929 | 5.89 | .844 | 11.36 | .930 | 3.59 |
| *MargExcIL (woMem)* | Liver | .967 | 5.89 | .965 | 14.99 | .962 | 17.45 | .965 | 6.32 | .972 | 4.72 | .948 | 4.66 | .979 | 1.83 |
| | Spleen | - | - | .963 | 1.21 | .956 | 2.42 | .963 | 1.17 | .957 | 12.39 | .885 | 23.77 | .971 | 1.19 |
| | Pancreas | - | - | - | - | .840 | 5.72 | .836 | 5.67 | .865 | 5.21 | .687 | 17.62 | .838 | 5.48 |
| | R Kidney | - | - | - | - | - | - | .968 | 1.63 | .945 | 5.44 | .870 | 6.94 | .937 | 3.71 |
| | L Kidney | - | - | - | - | - | - | .957 | 2.51 | .941 | 5.98 | .875 | 4.22 | .944 | 3.74 |
| | mean | **.967** | 5.89 | **.964** | 8.10 | .919 | 8.53 | .938 | **3.46** | **.936** | 6.75 | .853 | 11.44 | .934 | 3.19 |
| *MargExcIL (Ours)* | Liver | .967 | 2.83 | .966 | 3.41 | .966 | 7.05 | .965 | 7.98 | .971 | 3.27 | .947 | 4.85 | .978 | 1.86 |
| | Spleen | - | - | .962 | 1.18 | .962 | 1.21 | .962 | 1.30 | .956 | 4.60 | .888 | 6.27 | .969 | 1.26 |
| | Pancreas | - | - | - | - | .837 | 5.32 | .835 | 5.51 | .865 | 5.13 | .711 | 16.78 | .839 | 5.14 |
| | R Kidney | - | - | - | - | - | - | .968 | 1.40 | .946 | 6.14 | .846 | 7.44 | .943 | 3.48 |
| | L Kidney | - | - | - | - | - | - | .959 | 2.37 | .942 | 5.86 | .872 | 12.20 | .935 | 3.05 |
| | mean | **.967** | **2.83** | **.964** | **2.30** | .922 | 4.53 | .938 | 3.71 | **.936** | 5.00 | .853 | **9.51** | **.935** | **3.05** |
| *MargExc* [21] *(Upper Bound)* | Liver | .967 | 5.89 | .966 | 6.90 | .968 | 2.79 | .962 | 7.01 | .965 | 3.04 | .952 | 4.05 | .981 | 1.63 |
| | Spleen | - | - | .950 | 5.86 | .959 | 2.15 | .965 | 1.15 | .948 | 3.01 | .896 | 9.24 | .970 | 1.26 |
| | Pancreas | - | - | - | - | .841 | 5.92 | .848 | 4.83 | .862 | 5.67 | .677 | 20.41 | .849 | 4.94 |
| | R Kidney | - | - | - | - | - | - | .969 | 1.39 | .950 | 2.17 | .854 | 6.65 | .918 | 7.84 |
| | L Kidney | - | - | - | - | - | - | .965 | 3.96 | .943 | 2.28 | .898 | 10.64 | .935 | 6.39 |
| | mean | **.967** | 5.89 | .958 | 6.38 | **.923** | **3.62** | **.942** | 3.67 | .934 | **3.23** | **.855** | 10.20 | .931 | 4.41 |

and ILT, finding that old organs' logit is overwhelmed by the logit of background as the training stages progress. MiB [2] can get a good result compared with LwF and ILT because of remodeling background and foreground in training phase, thus avoiding the 'knowledge conflict' problem.

MargExc [21] also solves the 'knowledge conflict' problem, which is the most harmful factor in aggregating partially labeled dataset. Based on MargExc, our MargExcIL also performs well on DC and HD95, better than all other methods, *e.g.* MiB or the models trained separated for all organs, approaching upper bound result (DC: 0.938 vs 0.942 & HD95: 3.71 vs 3.67). In '*Model*$_{S3}$ *on Testing Datasets*' part in Table 3, MargExcIL even performs better than MargExc [21]. These results prove IL might have a practical potential in clinical scenario.

**Effectiveness of Memory Module.** In Table 3, we also show the results of the 4 intermediate stages of our method, in '$Model_{S*}$ *on Validation Sets*' part. '$(woMem)$' means our method without memory module and corresponding loss functions. '$_{swin}$' means that we modify encoder of our network designed by nnUNet to Swin Transformer [12], which can also assist in proving the effectiveness of our memory module. Without memory module, we can also obtain the same level performance in last stage, but it's *not stable in the middle stages*, e.g., liver's HD95 get worse dramatically in stage 2 and stage 3. This uncertainty factor in our IL system is not acceptable. We believe that this phenomenon is caused by the variation in the image distribution or field-of-view (FOV) in different datasets. Our memory module stores a prior knowledge of old class to stabilize the whole IL system. Compared with MargExc [21], we also achieve a comparable performance.

## 5 Conclusion

To unleash the potential from a collection of partially labeled datasets and to settle the efficiency, storage, and ethical issues in current methods, we introduce an incremental learning (IL) mechanism with a practical *four*-stage setting and verify the implementation potential of IL in MOS. IL methods have a natural adaptability to medical image scenarios due to the relatively fixed anatomical structure of human body. The introduced light memory module and loss functions can also stabilize the IL system in practice via constraining the representation of different categories in feature space. We believe that IL holds a great promise in addressing the challenges in real clinics.

## References

1. Bennett, L., et al.: 2015 miccai multi-atlas labeling beyond the cranial vault - workshop and challenge (2015). https://doi.org/10.7303/syn3193805
2. Cermelli, F., Mancini, M., Bulo, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9233–9242 (2020)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
4. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging **39**(11), 3619–3629 (2020)
5. Heller, N., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv:1904.00445 (2019)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

7. Jin, Z., et al.: Mining contextual information beyond image for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7231–7241 (2021)
8. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. **114**(13), 3521–3526 (2017)
9. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2935–2947 (2017)
10. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
11. Liu, Y., Schiele, B., Sun, Q.: Meta-aggregating networks for class-incremental learning. arXiv preprint arXiv:2010.05063 (2020)
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
13. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems 30 (2017)
14. Ma, J., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem. IEEE Trans. Pattern Anal. Mach. Intell., 1 (2021). https://doi.org/10.1109/TPAMI.2021.3100536
15. Marra, F., Saltori, C., Boato, G., Verdoliva, L.: Incremental learning for the detection and classification of gan-generated images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2019). https://doi.org/10.1109/WIFS47025.2019.9035099
16. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: the sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
17. Michieli, U., Zanuttigh, P.: Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
18. Ozdemir, F., Goksel, O.: Extending pretrained segmentation networks with additional anatomical structures. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1187–1195 (2019). https://doi.org/10.1007/s11548-019-01984-4
19. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2001–2010 (2017)
20. Rusu, A.A., et al.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
21. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Medical Image Analysis, p. 101979 (2021)
22. Shi, K., Bao, H., Ma, N.: Forward vehicle detection based on incremental learning and fast R-CNN. In: 2017 13th International Conference on Computational Intelligence and Security (CIS), pp. 73–76 (2017). https://doi.org/10.1109/CIS.2017.00024
23. Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3400–3409 (2017)
24. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv:1902.09063 (2019)

25. Tasar, O., Tarabalka, Y., Alliez, P.: Incremental learning for semantic segmentation of large-scale remote sensing data. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **12**(9), 3524–3537 (2019)
26. Xu, J., Zhu, Z.: Reinforced continual learning. Advances in Neural Information Processing Systems 31 (2018)
27. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1195–1204 (2021)
28. Zhou, S.K., et al.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE (2021)
29. Zhou, S.K., Rueckert, D., Fichtinger, G.: Handbook of Medical Image Computing and Computer Assisted Intervention. Academic Press (2019)
30. Zhou, Y., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10672–10681 (2019)