Low-Rank Continual Pyramid Vision Transformer: Incrementally Segment Whole-Body Organs in CT with Light-Weighted Adaptation

Vince Zhu[1,2], Zhanghexuan Ji[1], Dazhou Guo[1], Puyang Wang[3], Yingda Xia[1], Le Lu[1], Xianghua Ye[4], Wei Zhu[2], Dakai Jin[1]

[1] DAMO Academy, Alibaba Group
[2] State University of New York at Stony Brook, USA
[3] Johns Hopkins University, USA
[4] The First Affiliated Hospital Zhejiang University, Hangzhou, China
{zhanghexuan.ji, dakai.jin}@alibaba-inc.com

Abstract. Deep segmentation networks achieve high performance when trained on specific datasets. However, in clinical practice, it is often desirable that pretrained segmentation models can be dynamically extended to enable segmenting new organs without access to previous training datasets or without training from scratch. This would ensure a much more efficient model development and deployment paradigm accounting for the patient privacy and data storage issues. This clinically preferred process can be viewed as a continual semantic segmentation (CSS) problem. Previous CSS works would either experience catastrophic forgetting or lead to unaffordable memory costs as models expand. In this work, we propose a new continual whole-body organ segmentation model with light-weighted low-rank adaptation (LoRA). We first train and freeze a pyramid vision transformer (PVT) base segmentation model on the initial task, then continually add light-weighted trainable LoRA parameters to the frozen model for each new learning task. Through a holistically exploration of the architecture modification, we identify three most important layers (i.e., patch-embedding, multi-head attention and feed forward layers) that are critical in adapting to the new segmentation tasks, while retaining the majority of the pre-trained parameters fixed. Our proposed model continually segments new organs without catastrophic forgetting and meanwhile maintaining a low parameter increasing rate. Continually trained and tested on four datasets covering different body parts of a total of 121 organs, results show that our model achieves high segmentation accuracy, closely reaching the PVT and nnUNet upper bounds, and significantly outperforms other regularization-based CSS methods. When comparing to the leading architecture-based CSS method, our model has a substantial lower parameter increasing rate (16.7% versus 96.7%) while achieving comparable performance.

1 Introduction

Multi-organ and tumor segmentation, one of the most essential medical image analysis tasks, has been widely studied in the literature [5,19,20]. With the

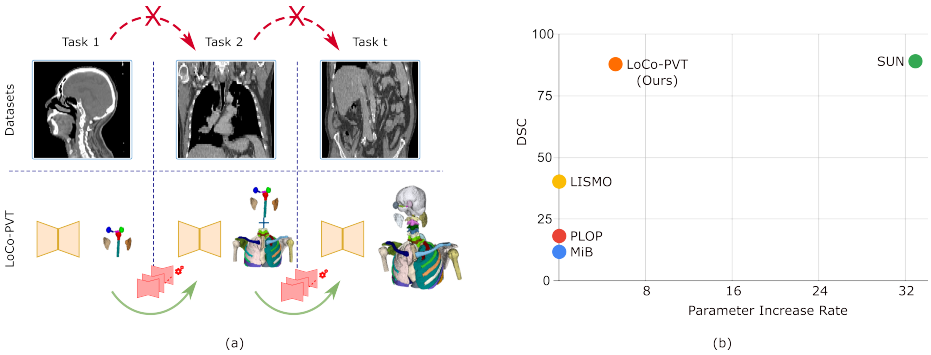

Fig. 1. Illustration of the continual multi-organ segmentation (a). At each continual learning step, only the previously trained model is available (green arrow). Previous datasets are not accessible. Illustration of the segmentation performance versus parameter increasing rate of continual multi-organ segmentation methods.

fast development in deep learning segmentation techniques, deep segmentation networks trained on specific datasets achieve high performance comparable to those of medical experts [8,10,11,17,22,26]. However, current deep segmentation approaches are not capable of updating the trained models effectively when new segmentation classes are incrementally added, although in clinical practice it is desirable that pre-trained segmentation models can be dynamically extended to segment new organs without access to previous training datasets. Illustrated in Fig. 1, this preferred process can be viewed as a continual semantic segmentation (CSS) problem, which is a non-trivial task because deep learning models suffer from catastrophic forgetting when fine-tuned directly on new dataset [12,15,18].

CSS is emerging very recently in the natural image domain [1,4,27], and the most common CSS approaches adopt the regularization constraint network training via knowledge distillation to reduce the forgetting of old knowledge while learning new classes. However, since entire network parameters are updated on the training of new classes, it is extremely difficult to achieve high performance on both old and new classes. CSS has been rarely studied in medical imaging field [9,14,16,28]. Ozdemir et al. uses only 9 patients and 2 organ labels to develop a regularization-based continual segmentation model [16]. Liu et al. adopts the MiB loss [1] and prototype matching to continually segment a small number of 5 abdominal organs [14], and Zhang et al. utilizes the pseudo-labels and clip-embedded controller head to segment 13 abdominal organs [28]. Note that [14,28] both only focus on a limited organs in the abdomen CT, and when involving a large number of organs of various body parts, such as in whole-body CT scans, they suffer severe performance degradation (as demonstrated in our experiments later). Recently, a new architecture-based CSS method [9] is proposed that avoids forgetting by freezing the CNN encoder after the initial task and sequentially adding separate decoder for each new task. Although it achieves high performance without forgetting, the method is less scalable because model

parameters increase dramatically as new tasks are added (see Figure 1(b)). The completely frozen encoder also lacks of extensibility [9].

In this work, we aim to develop a new CSS method that avoids the catastrophic forgetting and meanwhile circumvent the model parameter explosion issue in [9]. To achieve this goal, we take inspirations from two categories of recent technique advancements. First, vision transformer (ViT) is widely used in recent applications [3,13,23], and exhibits superiority in global feature extraction, self-supervised large model pretraining, and multi-modality learning as compared to the CNN-based models. In medical imaging, ViT-based models also demonstrate great potential for multi-organ segmentation task [21,25,29], since many of them exhibit comparable performance with the leading CNN-based models [8]. Considering the capacity advantage of ViT models and its flexibility in being extended to diverse tasks, we envision that ViT-based architecture is suitable for CSS task. Second, many recent parameter efficient fine-tuning (PEFT) methods are demonstrated to be effective when adapting the large scale pretrained language model to different downstream applications [2,6,7]. For example, low-rank adaptation (LoRA) [7] is one of the most popular and effective PEFT methods by freezing the pretrained model weights and injects trainable rank decomposition matrices into the linear or convolution layer of ViT. Hence, we assume that PEFT is capable of extending model's capacity to segment new organs with minor increased model parameters.

Motivated by the above observations, we propose a new architecture-based CSS method for continual whole-body organ segmentation using pyramid vision transformer with LoRA. We adopt the UniMISS [25] pretrained 3D pyramid vision transformer (PVT) as backbone due to its large scale medical image pretraining and the leading performance on downstream segmentation tasks. To circumvent the issue of catastrophic forgetting, we introduce a subset of trainable parameters for each new task. Unlike previous methods that append a bulky decoder for each task [9], our approach utilizes LoRA on selected PVT layers to incrementally expand its capacity for segmenting new organs. Following the original LoRA configuration, a group of LoRA matrices are first injected to query & value projection layers in multi-head attention to enhance the feature extraction. Furthermore, through a holistically exploration of the architecture modification, we inject LoRA matrices to the feed-forward network (FFN) to provide extra feature aggregation capability necessary for adapting to new unseen tasks. Additionally, we further extend LoRA matrices to 3D convolution in patch embedding layers of encoder and the last layer of decoder, which it critical to handle the large spacing variation in different medical segmentation tasks. Continually trained and tested on four datasets covering different body parts of a total of 121 organs, results show that our model achieves high segmentation accuracy, closely reaching the PVT [25] and nnUNet [8] upper bounds, and significantly outperforms other regularization or pseudo-label based CSS methods [1,4,28].

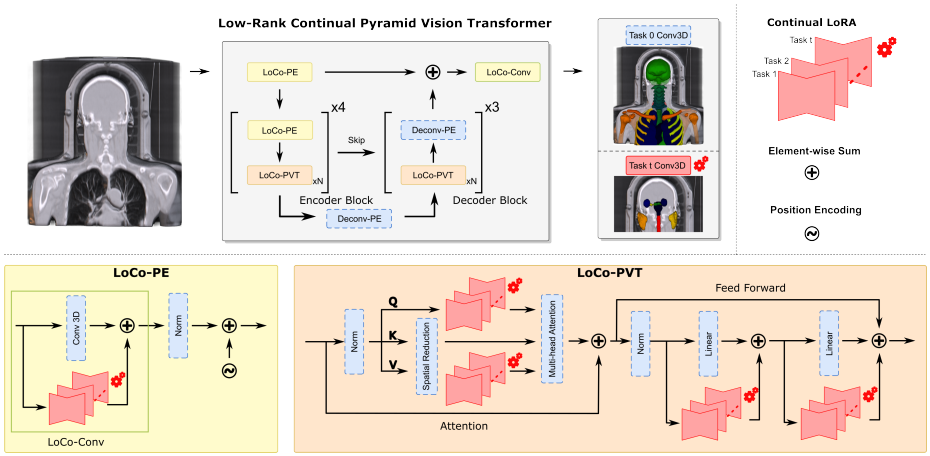

Fig. 2. Overall framework of the proposed low-rank continual pyramid vision transformer (LoCo-PVT) network for continual whole-body organ segmentation, which is composed of a stack of encoder and decoder blocks, where each block contains a patch embedding (PE) layer and multiple LoCo-PVT layers. Encoder PE layer (LoCo-PE) has a convolution layer with stride 2 for downsampling, while decoder PE layer (Deconv-PE) uses deconvolution layer for upsampling instead. Continual LoRA is added on linear layers for Q/V projection in multi-head attention and feed-forward network in LoCo-PVT, and is also added on convolution layers (LoCo-Conv) in LoCo-PE. The base network is frozen (colored in blue) after training the inital task 0. At each following continual learning step, a set of trainable LoRA parameters and a new segmentation output layer (colored in red) are added for new task adaptation.

2 Method

2.1 Problem Formulation

Figure 2 illustrates the proposed low-rank continual (LoCo) multi-organ segmentation framework. We adopt the UniMISS [25] pretrained 3D PVT as backbone. Subsequently, the 3D PVT undergoes further training with the TotalSegmentator dataset [24]. After this additional training, the PVT backbone, as depicted by the blue dashed-line blocks in Figure 2, is fixed throughout the subsequent training process. For the remaining tasks, let $D = \{D_1, \ldots, D_T\}$ represent the datasets sequence. The model is trained sequentially on each D_t where $t \in \{1, \ldots, T\}$ and will not re-access D_t after training is complete. Consider the t^{th} dataset $D_t = \{X_k^t, Y_k^t\}_{k=1}^{N_t}$ that compromises O_t organ classes, and assuming (X^t, Y^t) denote all input images and the corresponding segmentation masks in D_t, the prediction map for voxel location i and organ class c_j is given by

$$\hat{Y}^t(i) = f\left(Y^t(i) = c_j | X^t; W_0, W_{L^t}, W_{S^t}\right), \quad \hat{\mathbf{Y}} = \bigcup_{t=1}^{T} \hat{Y}^t, \tag{1}$$

where f denotes the transformer neural networks with frozen parameters W_0 trained on initial task and task-specific trainable LoRA parameters W_{L^t} where $L^t = (A_l^t, B_l^t)_{l=1}^{M^t}$ denotes the pairs of low rank matrices for the t^{th} dataset, and W_{S^t} denotes the trainable task-specific Sigmoid output. The final prediction $\hat{\mathbf{Y}}$ is the union of all previous predictions with possible class overlapping.

2.2 LoCo-PVT: LoRA Continual Vision Transformer Layer

Our LoCo-PVT framework inherits key advantages of both methods and is customized for continual multi-organ segmentation. The low-rank adaptations are enabled within every transformer block as well as the patch embedding modules of the encoder. For each dataset D_t, we associate a small set of trainable LoRA parameters $L^t = (A_l^t, B_l^t)_{l=1}^{M^t}$ where (A_l^t, B_l^t) denotes LoRA matrices of the l^{th} LoCo-PVT block and M^t represents the total number of trainable LoCo-PVT blocks for the t^{th} dataset.

LoCo-MHA & LoCo-FFN For a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times c}$ and the t^{th} dataset, we constrain its update by representing the latter with a low-rank decomposition $W_0 + \Delta W^t = W_0 + B^t A^t$, where $B^t \in \mathbb{R}^{d\times r}$, $A^t \in \mathbb{R}^{r\times c}$, and the rank $r \ll \min(d, c)$. During training, W_0 is frozen and does not receive gradient updates, while A^t, B^t contain trainable parameters. Both W_0 and ΔW are multiplied with the same input, and their respective output vectors are summed coordinate-wise. The forward pass of $h = W_0 x$ can summarized as

$$h = W_0 x + \Delta W^t x = W_0 x + B^t A^t x. \tag{2}$$

Each A_t is initialized with random Gaussian and B_t with a zero matrix, resulting in $\Delta W^t = B^t A^t$ being zero at the start of training. Then, $\Delta W^t x$ is scaled by $\frac{\alpha}{r}$ where α is a constant in r.

Although LoRA may be applied to any dense layer, Hu et al., [7] shows that applying it to queries and values of the MHA module yields the most significant performance gains. Therefore, we adopt a similar design choice in each LoCo-PVT block and uses a higher r to accommodate for the greater complexity in learning from visual signals.

2.3 LoCo-PE: LoRA Continual Patch Embedding Layer

The patch embedding modules project input patches into implicit embedding space of lower dimensions. Each encoder stage of LoCo-PVT is accompanied by a separate embedding module which extracts feature maps for various resolutions. Such spatial information are critical for training robust continual segmentation models across different datasets. Desirably, one should allocate trainable parameters to all convolutional projectors for each dataset. Compared to dense layers, it is observed that the inclusion of LoRA in convolutional layers resulted in a significant increase in number of parameters. To mitigate this issue, the application of convolutional LoRA is confined exclusively to the encoding PE layers,

i.e., LoCo-PE layer, which incorporates LoRA-enabled 3D convolutions for projecting input patches to the embedding space in each encoder stage. To enhance feature aggregation and projection from all scales for the segmentation output, LoRA is also enabled in the second to the last convolution layer in decoder.

In each 3D convolutional weight matrix $\delta_0 \in \mathbb{R}^{d\times c\times k^3}$ of dataset D_t, a pair of matrices (A^t, B^t) with the same rank r, where $B^t \in \mathbb{R}^{dk^2\times rk}$, $A^t \in \mathbb{R}^{rk\times ck}$, k is kernel size. The forward pass of the convolutional operations are

$$\delta_0 * x + \Delta\delta * x = \delta_0 * x + (B^t A^t) * x, \tag{3}$$

where δ_0 is the frozen convolution weights from the pretrained W_0.

Model Inference: To merge the output probability maps from all learned tasks, we follows the body-part-aware output merging method from SUN [9], which precomputes the average body part distribution map for each dataset, applies body-part regression over testing scans to eliminate the out-of-distribution body-part region from each task's prediction, then uses entropy-based ensemble to combine the prediction from all tasks. No task ID is required during inference.

3 Experiments and Results

Datasets: We evaluated the proposed model using the public dataset **TotalSegmentator** [24] (TotalSeg) as task 0 for base model training, which consists of 1204 CT scans of different body parts with 103 labeled anatomical structures (face label is removed). Similar to SUN [9] dataset setting, we conduct continual segmentation on three in-house datasets which cover chest body part, head-neck body part and an esophageal dataset with tumor. **Chest organ dataset** (CHO) contains 153 chest CT scans with 16 labeled chest organs, including 7 overlapping organs with TotalSeg and 9 new organs. **Head-neck organ dataset** (HNO) includes 244 head & neck CT scans with 9 new organs are annotated. The last **esophageal dataset** (EsoTumor) contains 567 CT scans of esophagus with tumor, which is more challenging for esophagus segmentation. We use 80% : 20% split for training and testing. At the final stage, a total of 121 organs are learned from all datasets. The median voxel resolutions are $1.5\times1.5\times1.5$mm, $1\times1\times2$mm, $0.7\times0.7\times5$mm, and $1\times1\times5$mm for TotalSeg, HNO, CHO and EsoTumor datasets, respectively.

CSS Protocols, Baselines and Metrics: In CSS experiments, the model is updated on a sequence of datasets. At each step, only the current dataset is used for training while all the previous datasets are not accessible. Following SUN [28] setting, two CSS orders are validated in order to demonstrate the robustness of the method. Order A: *TotalSeg* $\rightarrow$ *CHO* $\rightarrow$ *HNO* $\rightarrow$ *EsoTumor*. Order B: *TotalSeg* $\rightarrow$ *HNO* $\rightarrow$ *CHO* $\rightarrow$ *EsoTumor*, as shown in Table 1. We compare our method with 4 leading CSS works: 2 popular regularization-based baselines (MiB [1], PLOP [4]), a regularization-based method in medical (LISMO [14]) and an latest architecture-based method (SUN [9]). All methods are implemented in nnUNet data preprocessing and augmentation framework. Our method uses PVT

Table 1. Final step results of benchmark CSS methods on two orders of the selected multi-organ datasets. Dataset names are followed by their class numbers. Mean DSC (%, ↑) and HD95 (mm, ↓) are evaluated on each dataset as well as all classes (All). 'PIR (%, ↓)': parameter increasing rate of the final model (after three continual steps) compared to the size of base model trained on TotalSeg (initial step).

Methods	TotalSeg (103)		CHO (16)		HNO (9)		EsoTumor (1)		All (121)		PIR
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	
Order A: TotalSeg → CHO → HNO → EsoTumor											
MiB [1]	9.18	116.38	28.55	20.36	9.21	7.40	87.35	4.43	12.19	96.00	0
PLOP [4]	37.92	53.49	66.98	19.60	34.68	15.55	83.43	5.97	41.65	46.27	0
LISMO [14]	11.71	137.65	43.07	29.07	9.22	12.93	87.47	4.45	16.01	114.45	0
SUN [9]	91.93	3.44	84.33	5.20	84.79	2.67	86.59	5.15	90.45	3.62	96.7
Ours	91.07	4.09	81.51	5.77	82.90	3.08	84.56	5.48	89.26	4.24	16.7
Order B: TotalSeg → HNO → CHO → EsoTumor											
MiB [1]	11.24	145.78	78.65	7.23	9.09	24.83	87.27	4.36	20.04	119.06	0
PLOP [4]	31.58	63.78	79.47	7.09	22.78	11.09	83.19	6.04	37.31	52.63	0
LISMO [14]	15.01	90.54	79.49	7.36	8.93	9.13	87.32	4.31	23.14	73.88	0
SUN [9]	91.93	3.44	84.33	5.20	84.79	2.67	86.59	5.15	90.45	3.62	96.7
Ours	91.07	4.09	81.51	5.77	82.90	3.08	84.56	5.48	89.26	4.24	16.7
Single Task Upperbound											
PVT [25]	91.07	4.09	83.56	5.25	84.67	2.70	87.02	5.16	89.66	4.15	300
nnUNet [8]	91.93	3.44	84.48	5.14	84.95	2.66	87.62	4.39	90.49	3.60	300

as backbone, while the other 4 methods are based on CNN. In this study, the upper bound of both PVT and nnUNet on each dataset are listed in Table 1. We report the final continual segmentation performance using the Dice similarity coefficient (DSC) and the 95% Hausdorff distance (HD95).

Implementation Details: The 3D PVT base model is the same as UniMISS [25] model-small, which contains 4 encoder blocks with $[2, 3, 4, 3]$ PVT layers each and 3 decoder blocks with [3, 4, 3] PVT layers each. The stride is 2 for the convolution in encoder PE and deconvolution in decoder PE for down-/up-sampling purpose. For LoRA setting, we set rank as 64, 16 for query/value layers and FFN in LoCo-PVT and 16 for convolution layers in LoCo-PE/LoCo-Conv; LoRA alpha is consistently set as half of corresponding rank. The encoder is initialized from Unimiss self-supervised pretrained parameters. Following Unimiss setting, batch size of 2 and patch size of $224 \times 224 \times 32$ are used for all datasets and the experiments. The ratio between the training and validation set is 4:1. All experiments are trained using AdamW and we set 6000 epochs for training base model on TotalSeg and 500 epochs for continual training steps. The initial learning rate for PVT is set as $1e^{-4}$ and weight decay as $3e^{-5}$. Models are trained on single NVIDIA A100 GPU.

Comparisons to Other State-of-the-art Approaches: The final continual segmentation results on two orders and single task upper bounds are shown in Table 1. On the previously learned three datasets and all organs, our method significantly outperforms 3 regularization-based methods (MiB [1], PLOP [4], LISMO [14]) in both orders, where the severe catastrophic forgetting of these methods could be caused by large domain gap between different body parts. On the other hand, our method and SUN are both architecture-based methods

Table 2. One-step continual segmentation from TotalSeg to CHO trained with different LoRA ablation settings. LoRA ranks for Attn-QV, FFN and PE-Conv are set as 64, 16, 16, separately.

LoRA Settings	Attn-QV	FFN	PE-Conv	DSC	HD95
Base	✓			76.38	7.96
Base w. FFN	✓	✓		79.84	7.07
Base w. PE-Conv	✓		✓	80.36	6.30
Ours (Full)	✓	✓	✓	81.51	5.77

LoRA Settings	DSC	HD95
Encoder only	76.89	7.88
Decoder only	77.21	7.52
Ours (Full)	81.51	5.77

hence have no forgetting over past tasks and are order invariant when base training dataset is the same (TotalSeg). Although SUN has a slightly higher mean Dice of 90.45% than our 89.26% on all organs, our parameter increasing rate is significant lower than SUN (16.7% vs. 96.7% on 3 continual tasks), since SUN adds an entire decoder for each new task while our method adds light-weighted low-rank adaptors in selected layers, which only increases 5.56% per task. Note that, there is also a small gap between nnUNet upper bound 90.49% and PVT upper bound 89.66% on all organs, which shows a potential capability difference between nnUNet and PVT and might be the cause of the tiny performance gap between SUN (nnUNet-based) and LoCo-PVT (1.19%). Our proposed method closely reaches the PVT upper bound on all organs with a marginal 0.4% drop in DSC and a 0.9mm increase in HD95, which demonstrates the efficiency and effectiveness of continual LoRA with a well-trained frozen PVT.

Ablation Study: To demonstrate the importance of each continual LoRA components in the proposed LoCo-PVT network, we also conduct two ablation studies using one-step continual segmentation from TotalSeg to CHO, shown in Table 2. In the left table, various LoRA combinations are evaluated over three PVT components, including query & value projection layer in multi-head attention (Attn-QV), feed-forward network in transformer layer (FFN) and 3D convolution in patch embedding layer (PE-Conv). Compared to 'Base' setting, adding extra LoRA to FFN increases the CHO segmentation performance by 3.46%, from 76.38% to 79.84%; adding extra LoRA to PE-Conv results in 80.36%, which gains 3.98%; our full LoCo design with LoRA in all three components further boosts the performance to 81.51% and reduces HD95 from 7.96mm to 5.77mm. This ablation result shows that it is effective and essential to add LoRA in both FFN, which provides extra ability to project and ensemble new features from attention, and PE-Conv, which makes adaptation or localization on different patch resolution. In the right table, we further study the effect of adding LoRA in either encoder or decoder. In 'encoder only' setting with frozen decoder, the mean DSC drops to 76.89% and HD95 rises to 7.88mm; Similarly, in 'decoder only' setting with frozen encoder, the mean DSC reduces to 77.21% and HD95 increases to 7.52mm. The results shows that LoRA in both encoder (feature extraction) and decoder (organ localization) helps enhancing the adaptation ability of the network on new tasks and works equally important for LoCo-PVT network to get comparable performance with the upper bound. This ablation study validates the necessity of each component in our proposed network.

4 Conclusion

In this paper, we propose a new LoCo-PVT framework which combines LoRA with ViT for continual whole-body organ segmentation. We train and freeze a PVT base model on the initial task, then continually add light-weighted trainable LoRA parameters to the frozen base model, which avoids catastrophic forgetting and adapts the model to new tasks while maintaining a low parameter increasing rate. Our method achieves very high accuracy on four datasets covering different body parts, closely reaching the PVT upper bound, and outperforms other regularization-based methods. When comparing to leading architecture-based CSS method, our model exhibits a significantly lower parameter increase rate while achieving comparable performance. This efficiency highlights the effectiveness of our approach in optimizing resource use without compromising on the quality of organ segmentation. Future works include extending the LoCo-PVT to multi-modality datasets and other light-weighted ViT adaptation methods.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cermelli, F., Mancini, M., Bulo, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9233–9242 (2020)
2. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., et al.: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence **5**(3), 220–235 (2023)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4040–4050 (2021)
5. Guo, D., Jin, D., Zhu, Z., Ho, T.Y., Harrison, A.P., Chao, C.H., Xiao, J., Lu, L.: Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4223–4232 (2020)
6. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

9. Ji, Z., Guo, D., Wang, P., Yan, K., Lu, L., Xu, M., Wang, Q., Ge, J., Gao, M., Ye, X., et al.: Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21140–21151 (2023)
10. Jin, D., Guo, D., Ge, J., Ye, X., Lu, L.: Towards automated organs at risk and target volumes contouring: Defining precision radiation therapy in the modern era. Journal of the National Cancer Center **2**(4), 306–313 (2022)
11. Jin, D., Guo, D., Ho, T.Y., Harrison, A.P., Xiao, J., Tseng, C.K., Lu, L.: Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Medical Image Analysis **68**, 101909 (2021)
12. Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C.: Measuring catastrophic forgetting in neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
14. Liu, P., Wang, X., Fan, M., Pan, H., Yin, M., Zhu, X., et al.: Learning incrementally to segment multiple organs in a CT image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 714–724. Springer (2022)
15. Ma, C., Ji, Z., Huang, Z., Shen, Y., Gao, M., Xu, J.: Progressive voronoi diagram subdivision enables accurate data-free class-incremental learning. In: The Eleventh International Conference on Learning Representations (2023)
16. Ozdemir, F., Fuernstahl, P., Goksel, O.: Learn the new, keep the old: Extending pretrained models with new anatomy and images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 361–369. Springer (2018)
17. Raju, A., Ji, Z., Cheng, C.T., Cai, J., Huang, J., Xiao, J., et al.: User-guided domain adaptation for rapid annotation from user interactions: a study on pathological liver segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–467. Springer (2020)
18. Shen, Y., Ji, Z., Ma, C., Gao, M.: Continual domain adversarial adaptation via double-head discriminators. In: International Conference on Artificial Intelligence and Statistics. pp. 2584–2592. PMLR (2024)
19. Shi, F., Hu, W., Wu, J., Han, M., Wang, J., Zhang, W., Zhou, Q., Zhou, J., Wei, Y., Shao, Y., et al.: Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nature Communications **13**(1), 1–13 (2022)
20. Tang, H., Chen, X., Liu, Y., Lu, Z., You, J., Yang, M., Yao, S., Zhao, G., Xu, Y., Chen, T., et al.: Clinically applicable deep learning framework for organs at risk delineation in ct images. Nature Machine Intelligence **1**(10), 480–491 (2019)
21. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
22. Wang, P., Guo, D., Zheng, D., Zhang, M., Yu, H., Sun, X., Ge, J., Gu, Y., Lu, L., Ye, X., et al.: Accurate airway tree segmentation in ct scans via anatomy-aware multi-class segmentation and topology-guided iterative learning. IEEE Transactions on Medical Imaging (2024)

23. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
24. Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M.: Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. arXiv preprint arXiv:2208.05868 (2022)
25. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)
26. Ye, X., Guo, D., Ge, J., Yan, S., Xin, Y., Song, Y., Yan, Y., Huang, B.s., et al.: Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. Nature communications **13**(1), 1–15 (2022)
27. Zhang, C.B., Xiao, J.W., Liu, X., Chen, Y.C., Cheng, M.M.: Representation compensation networks for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7053–7064 (2022)
28. Zhang, Y., Li, X., Chen, H., Yuille, A.L., Liu, Y., Zhou, Z.: Continual learning for abdominal multi-organ and tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 35–45. Springer (2023)
29. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)

Low-Rank Continual Pyramid Vision Transformer: Incrementally Segment Whole-Body Organs in CT with Light-Weighted Adaptation — Supplementary Material

Table S1. Labeled organ list of each dataset. Please refer to TotalSegmentator[23] for detailed organ list of each group.

Datasets	**Organs**
TotalSeg	26 major organs, 8 vessels, 10 muscles, 59 bone instances
CHO	**overlapping**: aorta, heart, left/right lung, inferior vena cava, esophagus, pulmonary artery; **new**: airway, sternum, left/right thyroid, superior vena cava, pulmonary vein, scalenus muscle, scalenus anterior muscle, sternocleidomastoid muscle
HNO	brain stem, spinal cord, left/right eye, left/right optic nerve, chiasm, left/right parotid
EsoTumor	**overlapping**: esophagus with tumor

Table S2. Step-wise mean DSC (%, ↑) and forgetting rate (FR) (%, ↓) of CSS methods on each dataset and all 121 organs ('All'). EsoTumor is not included since no forgetting rate is calculated for the last continual step.

Methods	Step	TotalSeg		CHO		HNO		All	
		DSC	FR	DSC	FR	DSC	FR	DSC	FR
Order A: TotalSeg → CHO → HNO → EsoTumor									
MiB[1]	2	43.07	53.15	84.45	—	—	—	48.63	53.15
	3	12.53	86.37	39.79	52.88	84.62	—	21.01	69.63
	4	9.18	90.01	28.55	66.19	9.21	89.12	12.19	81.77
PLOP[4]	2	58.98	35.84	82.66	—	—	—	62.16	35.84
	3	40.11	56.37	70.31	14.94	82.88	—	46.89	35.65
	4	37.92	58.75	66.98	18.97	34.68	58.16	41.65	45.29
LISMO[14]	2	50.43	45.14	84.46	—	—	—	55.01	45.14
	3	15.02	83.66	44.24	47.62	84.9	—	23.59	65.64
	4	11.71	87.26	43.07	49.01	9.22	89.14	16.01	75.14
SUN[9]	2	91.93	0	84.33	—	—	—	90.91	0
	3	91.93	0	84.33	0	84.79	—	90.48	0
	4	91.93	0	84.33	0	84.79	0	90.45	0
Ours	2	91.07	0	81.51	—	—	—	89.78	0
	3	91.07	0	81.51	0	82.9	—	89.30	0
	4	91.07	0	81.51	0	82.9	0	89.26	0
Order B: TotalSeg → HNO → CHO → EsoTumor									
MiB[1]	2	23.36	74.59	—	—	84.87	—	31.63	74.59
	3	12.08	86.86	84.43	—	9.88	88.36	20.97	86.86
	4	11.24	87.77	78.65	6.85	9.09	89.29	20.04	61.30
PLOP[4]	2	47.32	48.53	—	—	83.71	—	52.21	48.53
	3	32.65	64.48	84.15	—	25.54	69.49	38.59	64.48
	4	31.58	65.65	79.47	5.56	22.78	72.79	37.31	48.00
LISMO[14]	2	27.21	70.40	—	—	84.94	—	34.97	70.40
	3	15.94	82.66	84.49	—	10.13	88.07	24.10	82.66
	4	15.01	83.67	79.49	5.92	8.93	89.49	23.14	59.69
SUN[9]	2	91.93	0	—	—	84.79	—	90.97	0
	3	91.93	0	84.33	—	84.79	0	90.48	0
	4	91.93	0	84.33	0	84.79	0	90.45	0
Ours	2	91.07	0	—	—	82.9	—	89.97	0
	3	91.07	0	81.51	—	82.9	0	89.30	0
	4	91.07	0	81.51	0	82.9	0	89.26	0