Locate, Crop and Segment: Efficient abdominal CT image segmentation on CPU

Yinyin Luo[†], Yue Liu[†], Wenbin Liu, Jingheng Dai, Xunliang Xiao, and Gang Fang[*]

Institute of Computing Science and Technology, Guangzhou University, Guangzhou, 510006, China
[†]Co-first authors
[*]Corresponding author
gangf@gzhu.edu.cn

Abstract. Although current deep learning based models have achieved tremendous successes in medical segmentation tasks, the deployment of such models on CPU only devices is still challenging due to the substantial computational resources required for segmentation inference, especially for 3D medical images. Small sized models capable of efficient inference have been proposed to mitigate the computational overheads, however these small models usually largely sacrifice the segmentation accuracy. In order to tackle the challenge in compliance with the requirements of MICCAI FLARE 2024 Challenge Task 2, i.e. deploying advanced 3D abdominal CT segmentation models in non-GPU environments while maintaining high accuracy, we introduce a multi-scale knowledge distillation method to train a student model that maximally retains the segmentation performance of the teacher model. In order to improve the segmentation performance of tiny organs and Overcome the quality issues of pseudo-labels themselves, We also design a weighted composite loss function to train the model. Furthermore, for efficient segmentation inference on CPU only devices, we introduce a liver-based Z-axis Region-of-Interest (RoI) localization strategy which effectively improve the segmentation efficiency. Experiments on the MICCAI FLARE 2024 datasets have shown significant improvements in both segmentation accuracy and efficiency. The proposed method achieves an average organ Dice Similarity Coefficient (DSC) of 88.70% and a Normalized Surface Dice (NSD) of 94.29% on the public validation set. In the FLARE 2024 Task2 online validation, the method achieved an average organ Dice Similarity Coefficient (DSC) of 88.47% and a Normalized Surface Dice (NSD) of 94.71%, with an impressive average inference time of 12.33 seconds. The code is available at https://github.com/lay-john/FLARE24-Task2.

Keywords: Semi-supervised · Deep learning · Organ segmentation · Knowledge distillation.

1 Introduction

The automatic and accurate semantic segmentation of medical images is a fundamental problem in medical image analysis, which serves as a crucial step in computer-aided diagnosis, surgical navigation, visual enhancement, radiotherapy, and biomarker measurement systems[27]. Segmentation accuracy is the most important factor to be considered when developing segmentation applications to aid medical image analysis, as it directly affects the patient diagnosis and the efficiency of clinical workflows. Based on this consensus, various methods, e.g. algorithm based methods and deep learning based methods, have been proposed to tackle this problem in recent years. For instance, U-Net[23] has shown superior segmentation accuracy by effectively utilizing skip connections to preserve spatial information. The DeepLab[2] series enhances multi-scale context and detail precision through the introduction of atrous convolution and Conditional Random Fields (CRFs). nnU-Net[12] offers an adaptive framework that optimizes performance for various datasets, showcasing remarkable versatility and effectiveness. Attention U-Net[22] leverages attention mechanisms to focus on crucial features, further improving segmentation accuracy. These methods have all achieved excellent segmentation results in different applications, advancing the development of medical image analysis. Among these methods, it is highly noted that nnU-Net has consistently demonstrated outstanding performance in different medical image segmentation tasks.

However, most current research works, including nnU-Net, neglect the importance of the computational efficiency of segmentation inference, i.e. the inference time and computational resource required, especially for 3D medical images. For an example, for a complete semantic segmentation of a 3D abdominal Computerized tomography (CT) image using nnU-Net, it would approximately take 60 seconds and 8G GPU memory, which is unbearable in real diagnosis practice. In fact, the memory consumption and GPU usage of these methods have led to a huge demand for computing resources, posing considerable challenges to the industrial deployment of the method. To mitigate the computational overheads, small sized models capable of efficient inference have been proposed, for an example EfficientNet[26], however previous researchers find that using smaller models often entails a compromise in performance[10]. Thus, this limitation prompts a critical challenge: how can we deploy state-of-the-art abdominal segmentation models in non-GPU environments without compromising segmentation accuracy.

In order to tackle the challenge in compliance with the requirements of MICCAI FLARE 2024 Challenge Task 2, i.e. deploying advanced 3D abdominal CT segmentation models in non-GPU environments while maintaining high accuracy, we introduce knowledge distillation and a pre-processing strategy in this paper. Specifically, in MICCAI FLARE 2024 Challenge Task 2, it is required to perform semantic segmentation of abdominal organ CT images using a CPU-based algorithm on a notebook computer with an 8GB memory limit. The task, which uses the same dataset as FLARE2022[15], involves segmenting 13 organs from CT images provided by over 20 medical groups, the organ labels are illustrated in Fig. 1. The dataset includes 2050 cases for model training, 250 cases

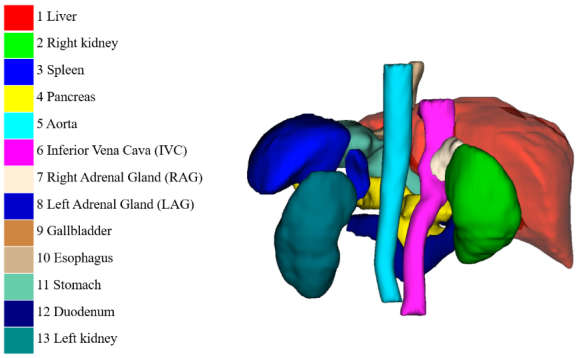

Fig. 1. Semantic labels of the 13 abdominal organs in FLARE 2024 Task2.

for validation, and 300 new, hidden test cases for final evaluation. The evaluation metrics include Dice Similarity Coefficient (DSC), Normalized Surface Dice (NSD), and seconds for inferring a single CT image(runtime).

Follow prior works, we also take advantage of small sized model for efficient inference, and we propose a multi-scale knowledge distillation method[31] to train a student model that maximally retains the segmentation performance of the teacher model. Knowledge distillation[9] is a technique where knowledge from a larger, more complex model (i.e. the teacher model) is transferred to a smaller, more efficient model (i.e. the student model), allowing the student model to achieve performance comparable to that of the teacher model while significantly reducing computational resources. Multi-scale knowledge distillation is commonly adopted in visual tasks to guide the training of a smaller student model by extracting multi-scale visual knowledge from a larger teacher model.Specifically, we perform multi-scale knowledge distillation by enforcing the student model to learn feature maps from the two intermediate layers of the teacher model by optimizing Mean Squared Error (MSE). In order to improve the segmentation performance of tiny organs and Overcome the quality issues of pseudo-labels themselves. we also propose a weighted composite loss algorithm based on DiceLoss and CELoss.

Furthermore, we introduce a pre-processing strategy to speed up the inference process. Given a 3D abdominal CT scan that is quite large in vertical size, the region of interest (RoI) for abdominal semantic segmentation may only occupy a portion of the scan, an example is shown in Fig. 2. It is unnecessary to perform inference on irrelevant areas. Thus, identifying and cropping the RoI is an effective strategy to reduce the inference time and the resource consumption. To address this, we propose a liver-based Z-axis RoI localization strategy to focus segmentation inference on the abdominal regions of interest. Previously, the approach for determining the RoI involved training a dedicated model specifically for RoI detection. This kind of methods induce increases in computational

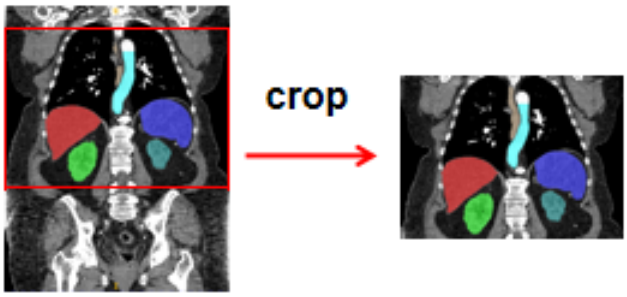

Fig. 2. The RoI only occupies a portion of the 3D abdominal CT scan.

complexity and time. Additionally, the RoI detection model often exhibits inferior performance compared to the segmentation model. Consequently, errors from the RoI detection model would propagate and adversely affect the final segmentation results. In contrast, our proposed strategy eliminates the need for a separate ROI detection model by leveraging our segmentation model to perform both ROI localization and segmentation. Our strategy is based on the accuracy of liver segmentation: as long as the Dice score for liver segmentation exceeds 60%, which is relatively easy to achieve, we can determine the ROI along the z-axis of the CT scan. This strategy minimizes the impact of poor input data quality and avoids the issues associated with a separate ROI detection model.

To summarize, in this paper, we propose a multi-scale knowledge distillation method and a Z-axis RoI localization strategy to tackle the MICCAR FLARE 2024 Challenge Task 2, the main contributions are summarized as follows: (1) Through multi-scale knowledge distillation, an efficient student model effectively absorbs the knowledge from the teacher model, thus maximally retains the performance of the teacher model while significantly reducing the computational overheads. (2) In order to improve the segmentation performance of tiny organs and Overcome the quality issues of pseudo-labels themselves, we also propose a weighted composite loss algorithm based on DiceLoss and CELoss. (3) To improve the speed of the segmentation inference, we further introduce a liver-based z-axis ROI localization strategy. Experimental results demonstrate the effectiveness of the proposed method in enhancing the performance of small sized model and accelerating the segmentation process.

2 Method

2.1 Preprocessing

- **Resample and normalization:** We resample the pixel spacing to (2.2838, 1.8709, 1.8709) for all cases, and clip the pixel value based on the Hounsfield units to [-160, 240], and normalize all the cases in [0, 1] to ensure data stability and consistency.
- **Data augmentation:** In order to prevent the model from over-fitting, data augmentation is used in this study. The augmentation approaches of nnU-Net methodology have been utilized.

2.2 Network Architecture

Here we define two nnUNet structures of different model size: nnU-Net-Teacher and nnU-Net-Student. nnU-Net-Teacher is constructed as a conventional nnU-Net structure for medical image segmentation tasks. nnU-Net-Student is constructed by modifying a conventional nnU-Net architecture by changing the initial number of channels from 32 to 16 and adding an additional convolutional layer with a channel count doubled while keeping the resolution unchanged. The network architecture of nnU-Net-Teacher and nnU-Net-Student are shown in Fig. 3. The model hyper-parameters and the input patch size of [80, 160, 160] are chosen to satisfy the memory requirement by the FLARE 2024 Task2 competition.

2.3 Proposed Method

Specifically, this paper proposes a multi-scale knowledge distillation method that initially trains a large Teacher model using the provided FLARE22 dataset and pseudo labels. This Teacher model is subsequently utilized to perform multi-scale knowledge distillation on a smaller Student model. A distillation loss (MSE) is calculated at the lowest two layers of the encoders in both the Teacher model and the Student model, and this distillation loss is incorporated into the Student loss to enhance the performance of the Student model through multi-scale knowledge distillation.When training the model. we also proposed a weighted compound loss algorithm based on DiceLoss and CELoss.Another contribution is the liver-based z-axis ROI localization technique. This technique involves scanning from the bottom to the top along the z-axis and stopping when a certain number of liver voxels are detected, using this patch to determine the z-axis ROI.

Multi-scale Knowledge Distillation: Initially, a large nnUNet model with an initial channel number of 32 is trained using Dice and Cross-Entropy loss for segmentation, achieving satisfactory performance. Subsequently, a smaller model is designed. To ensure consistency in the region where knowledge distillation is performed, an additional layer which does not change the resolution is added to the Student model, which has an initial channel number of 16, as illustrated in Fig. 3. During the training of the Student model, distillation losses are introduced

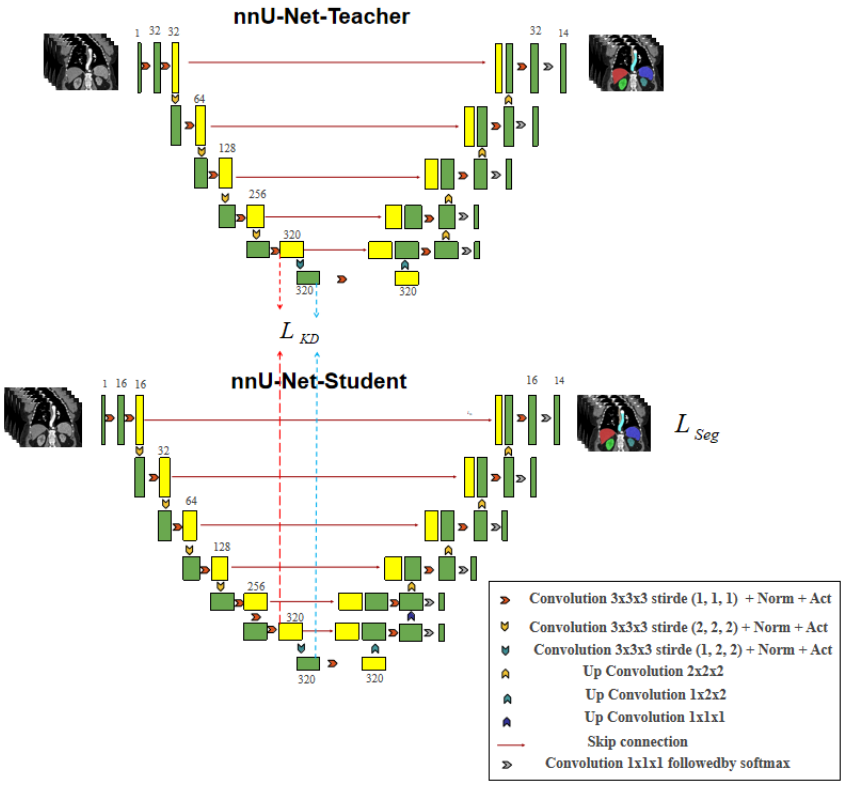

Fig. 3. Overview of the proposed method. (see Sect. 2.3 for details).

At the bottom two layers of the encoder. Both scales utilize MSE loss, and the distillation losses from these two scales are simply summed to obtain the total distillation loss. This total distillation loss is then multiplied by a weight w and added to the final loss.Here w is set to 0.5.Teacher model and Student model both use a deep supervision strategy.During the training phases of the two models, we use pseudo-labels generated from 2,000 unlabeled cases based on the winning solution[11] from FLARE 2022 , without any further processing.

Loss Function: The loss function combines weighted DiceLoss and CELoss, where CombLoss[25] converges significantly faster than cross-entropy loss. Specifically, different weights are applied to DiceLoss and CELoss. For the weight w1 of DiceLoss, we use the Mean DSC for each organ as reported in the winning solution of FLARE 2022[11]. For the weight w2 of CELoss, we allocate weights based on the size ratio of each organ, with smaller organs receiving higher weights. This approach ensures that the model focuses more on the quality of pseudo-labels and effectively considers the impact of organ size on the model.When train the Student model we need to add the total distillation loss.The loss function im-

proves the performance of the model. See Section 4.1 for details.The formula is as follows:

$$L_{Dice}(y, \hat{y}, w1) = \sum_{i}^{c_0} w1_i \left(1 - \frac{2\sum_{j=1}^{N} y_j^i \hat{y}_j^i}{\sum_{j=1}^{N} y_j^i + \hat{y}_j^i}\right) \tag{1}$$

$$L_{CE}(y, \hat{y}, w2) = \sum_{i}^{c_0} w2_i \left(-\frac{1}{N}\sum_{j=1}^{N}\left(y_j^i \log(\hat{y_j^i}) + (1 - y_j^i)\log(1 - \hat{y_j^i})\right)\right) \tag{2}$$

$$L_{KD} = w3 \cdot MSE_1 + w4 \cdot MSE_2 \tag{3}$$

$$L_{Seg} = \alpha_{dc} \cdot L_{Dice}(y, \hat{y}, w1) + \alpha_{ce} \cdot L_{CE}(y, \hat{y}, w2) + \alpha_{KD} \cdot L_{KD} \tag{4}$$

Where the y and $\hat{y}$ mean the ground truth and the predicted probability, respectively, and N is the number of pixels. α_{dc} and α_{ce} are the hyperparameters to balance the contribution of DiceLoss and CELoss. α_{dc} and α_{ce} are set to 0.5 in this study. α_{KD} is the hyperparameter of knowledge distillation. In this study, it is set to 0.5. $w3$ and $w4$ are hyperparameters on two scales of multi-scale knowledge distillation, and they are both set to 0.5 in this study.

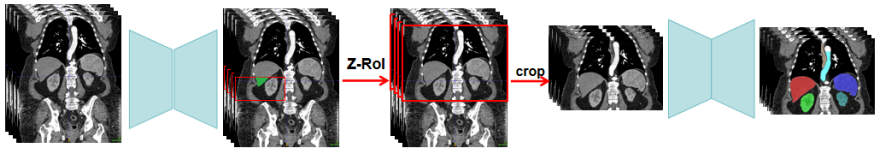

Fig. 4. The proposed liver-based Z-axis RoI localization strategy. (see Sect. 2.3 for details.)

Liver-based Z-axis RoI Localization: Prior to inference, the z-axis ROI is determined and the region is cropped for final inference, as shown in Fig. 4, we also wrote pseudocode for this process, as shown in Algorithm 1. Specifically, fixed sliding window regions are set for the x-axis and y-axis, both at $(0, 160)$, with a step size of 80 for sliding only along the z-axis starting from 0. A liver threshold, set to 20,000 in this study, is defined. During the z-axis sliding process, if the number of liver label voxels in the window is less than the threshold, the window slides to the next position; otherwise, the sliding stops. At this point,

the area of the window on the z-axis is $(i, i + \text{patch_size}[0])$, where i is the value of the lowest point of the z-axis window. Within this sliding window, the bounding box of the liver label is extracted, and based on the lower limit of the z-axis of this bounding box (minzidx), the final z-axis ROI range is $(i + \text{minzidx} - \frac{7}{8} \times \text{patch_size}[0],\ i + \text{minzidx} + 2 \times \frac{7}{8} \times \text{patch_size}[0])$. This provides a z-axis ROI region, which is then used for inference. The z-axis ROI region is defined as covering three z-axis sliding regions because with our patch_size $= (80, 160, 160)$, spacing $= (2.2838, 1.8709, 1.8709)$, and stride $= \left(\frac{7}{8}, \frac{7}{8}, \frac{7}{8}\right)$, all z-axis abdominal organ regions can be covered within three sliding windows.

Algorithm 1 Pseudocode for axis-based z-axis RoI positioning.

Input: CT scan, liver threshold (20000), patch size (80, 160, 160), spacing (2.2838, 1.8709, 1.8709), stride (7/8, 7/8, 7/8)
Output: z-axis ROI region

```
Initialize: x_range, y_range, z_start, step_size, liver_threshold, patch_size_z
Define sliding window regions for x and y axes:
x_window = x_range
y_window = y_range
Slide along the z-axis:
for i = z_start to end_of_z_axis step step_size do
    current_window = CT_scan[x_window, y_window, i : i + patch_size_z]
    liver_voxel_count = count_liver_voxels(current_window)
    if liver_voxel_count ≥ liver_threshold then
        break
    end if
end for
Determine z-axis ROI range:
z_lower_bound = i
Extract liver label bounding box within the sliding window:
liver_bbox = extract_liver_bbox(current_window)
minzidx = liver_bbox.z_lower_limit
Calculate final z-axis ROI range:
final_z_lower = z_lower_bound + minzidx − (7/8 × patch_size_z)
final_z_upper = z_lower_bound + minzidx + (2 × 7/8 × patch_size_z)
z_axis_ROI = (final_z_lower, final_z_upper)
return z_axis_ROI
```

2.4 Post-processing

In the post-processing stage, to save time, no post-processing operations were performed.

3 Experiments

3.1 Dataset and evaluation measures

The dataset is curated from more than 40 medical centers under the license permission, including TCIA [3], LiTS [1], MSD [24], KiTS [7,8], autoPET [6,5], AMOS [14], AbdomenCT-1K [21], TotalSegmentator [28], and past FLARE challenges [18,19,20]. The training set includes 2050 abdomen CT scans where 50 CT scans with complete labels and 2000 CT scans without labels. The validation and testing sets include 250 and 300 CT scans, respectively. The annotation process used ITK-SNAP [30], nnU-Net [13], MedSAM [16], and Slicer Plugins [4,17].

The evaluation metrics encompass two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—alongside one efficiency measures—runtime. These metrics collectively contribute to the ranking computation. During inference, GPU is not available where the algorithm can only rely on CPU.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements.

System	Ubuntu 22.04 LTS or Windows 10
CPU	Intel(R) Core(TM) i9-10900X CPU@3.70GHz
RAM	16×4×32GB; 2933MT/s
Programming language	Python 3.9.16
Deep learning framework	torch 2.1.0, torchvision 0.16.0
Specific dependencies	nnU-Net 1.7.0
Code	https://github.com/lay-john/FLARE24-Task2

Training protocols During the training phase, we set the batch size to 2 and randomly select all samples within each epoch. For each sample, we perform random patch cropping with patch sizes of (80, 160, 160). As for the optimizer, we utilize AdamW with a learning rate of 1e-2 and a weight decay of 1e-5. The learning rate updating follows the default mechanism of AdamW. Additional details are presented in Table 2.

Table 2. Training protocols for Teacher model.

Network initialization	
Batch size	2
Patch size	80×160×160
Total epochs	500
Optimizer	AdamW with weight decay(μ = 1e -5)
Initial learning rate (lr)	0.01
Lr decay schedule	halved by 200 epochs
Training time	35 hours
Loss function	DiceLoss and CELoss
Number of Teacher model parameters	30.8M[1]
Number of Teacher model flops	469.2262 G[2]
Teacher model CO_2eq	1.61908 Kg[3]
Number of Student model parameters	30.3M[4]
Number of Student model flops	128.1024 G[5]
Student model CO_2eq	0.73306 Kg[6]

4 Results and discussion

4.1 Quantitative results on validation set

Quantitative evaluation results are shown in Table 3, demonstrating that the proposed method achieves very promising results for major organs such as the liver, spleen, kidneys, and stomach. However, segmenting smaller organs remains highly challenging and requires further attention, particularly for very small organs with unclear boundaries, such as the adrenal glands and duodenum.

Table 3. Quantitative evaluation results.

Target	Public Validation		Online Validation	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)
Liver	97.09	98.41	97.14	98.12
Right Kidney	92.37	94.26	93.87	95.71
Spleen	96.82	98.89	96.11	98.07
Pancreas	89.25	97.80	86.13	96.50
Aorta	94.50	98.48	94.78	98.85
Inferior vena cava	89.27	91.02	89.48	92.20
Right adrenal gland	79.58	93.01	81.71	95.63
Left adrenal gland	79.64	91.86	81.48	94.57
Gallbladder	84.58	86.06	83.03	84.86
Esophagus	85.77	95.54	81.25	92.35
Stomach	91.38	94.41	91.96	95.92
Duodenum	81.68	94.41	79.65	92.79
Left kidney	91.18	92.66	93.53	95.72
Average	88.70	94.29	88.47	94.71

Tables 4, 5 and 6 present the final testing results of our proposed method in the FLARE 2024 Task2 across the Asian, European, and North American datasets, respectively. These tables list the performance metrics of our method, including the Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD), and inference time. Each metric is reported with both the mean and standard deviation (Mean ± Std), as well as the median along with the first and third quartiles (Median (Q1, Q3)).

Table 4. Final testing results of the proposed method on the FLARE 2024 Task2 Asian datasets.

Metric	Mean ± Std	Median (Q1, Q3)
DSC (%)	85.2 ± 6.2	87.8 (80.7, 90.5)
NSD (%)	92.1 ± 5.9	94.3 (88.3, 96.8)
Inference Time (s)	15.5 ± 3.7	13.7 (12.8, 18)

Table 5. Final testing results of the proposed method on the FLARE 2024 Task2 European datasets.

Metric	Mean ± Std	Median (Q1, Q3)
DSC (%)	87.4 ± 6.2	89.7 (84.7, 91.9)
NSD (%)	93.4 ± 6.1	95.7 (91.3, 98)
Inference Time (s)	16.4 ± 4.6	17.5 (12.8, 18.3)

Table 6. Final testing results of the proposed method on the FLARE 2024 Task2 North American datasets.

Metric	Mean ± Std	Median (Q1, Q3)
DSC (%)	87.6 ± 4.5	89.1 (85.4, 91)
NSD (%)	93.1 ± 4.7	94.7 (91, 96.7)
Inference Time (s)	12.6 ± 2.7	13 (12.7, 13.3)

To conduct a more comprehensive ablation study of our proposed method, we performed quantitative experiments, as shown in Table 7.Both of our proposed methods have shown improvements.

Table 7. Overview of Ablation Experiment Results. Proposed: Base+KD+Z-RoI.

Target	Base		Base+ KD		Base+KD+Z-RoI	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	96.59	97.97	97.07	98.25	97.09	98.41
Right Kidney	89.97	91.10	90.92	92.57	92.37	94.26
Spleen	96.68	98.78	96.81	98.86	96.82	98.89
Pancreas	88.07	97.41	88.91	97.43	89.25	97.80
Aorta	93.84	98.13	94.56	98.60	94.50	98.48
Inferior vena cava	85.93	87.43	88.83	90.58	89.27	91.02
Right adrenal gland	76.69	90.65	79.79	93.22	79.58	93.01
Left adrenal gland	74.77	87.59	76.05	88.34	79.64	91.86
Gallbladder	77.45	79.64	84.96	86.70	84.58	86.06
Esophagus	82.58	92.25	84.60	94.27	85.77	95.54
Stomach	88.93	92.87	91.12	94.75	91.38	94.41
Duodenum	80.26	92.10	81.09	93.66	81.68	94.41
Left kidney	86.89	89.45	88.01	89.60	91.18	92.66
Average	86.05	91.95	87.90	93.60	88.70	94.29

To visually demonstrate the impact of our method on inference speed on cpu, we conducted quantitative experiments on inference speed, as shown in Table 8.the length of the step is [7/8, 7/8, 7/8] times the window width for each axis.

Table 8. Overview of Ablation Experiment Results on inference speed. Proposed: Small nnU-Net + Z-RoI. Time is measured in seconds.

Target	Average time(s)
Big nnU-Net	38.21
Small nnU-Net	17.88
Small nnU-Net + Z-RoI	12.33

4.2 Qualitative results on validation set

In this section, we show the two good segmentation cases and two bad segmentation case, along with the time consumption for inference on several large CT scans.

Good segmentation cases: Fig. 5 presents examples of good segmentation results. In FLARETs_0001, both the Baseline and our proposed method demonstrate strong segmentation performance. In FLARETs_0012, while the overall segmentation performance remains satisfactory, our method exhibits a minor shortcoming where a portion of the right kidney is erroneously excluded and classified as background. This discrepancy may stem from our model's incorrect learning from the teacher model, leading to reduced sensitivity in certain localized regions, thereby failing to capture this specific area.

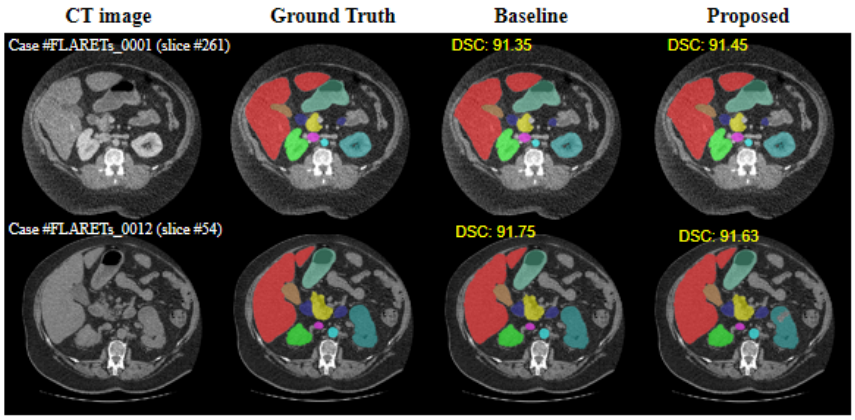

Fig. 5. Good segmentation cases from 50 validation set.

Bad segmentation cases: Fig. 6 presents examples of bad segmentation results.In FLARETs_0033, the baseline model demonstrates suboptimal performance, occasionally segmenting a single organ as two distinct organs. In contrast, our proposed method generally performs well; however, it misclassifies a portion of the duodenum as the stomach and exhibits poor segmentation performance for the pancreas. This is likely due to the anatomical proximity of the stomach and duodenum, especially at the junction between the lower end of the stomach and the beginning of the duodenum, where the boundary between the two organs is not always clearly defined. Additionally, the segmentation performance of the pancreas was suboptimal. This may because the pancreas is located deep within the abdomen, surrounded by other organs such as the stomach, duodenum, and liver, and has a long and variable shape. These anatomical features make it difficult to distinguish the pancreas from other tissues in CT images. In

FLARETs_0049, Both Baseline and our method have poor segmentation of the liver and inferior vena cava. As can be seen from the figure, there are patches of black areas in the liver of the FLARETs_0049 use case. These may be some diseased areas, resulting in poor model performance. For the inferior vena cava, the anatomical location of the inferior vena cava is close to many other important organs and blood vessels. The morphology may not be significantly different from the surrounding tissue, making it difficult for the model to accurately distinguish the inferior vena cava from the inferior vena cava and the shape and size of the inferior vena cava may vary greatly between different patients. This variation may increase the difficulty of model segmentation, which is difficult for our model to learn.

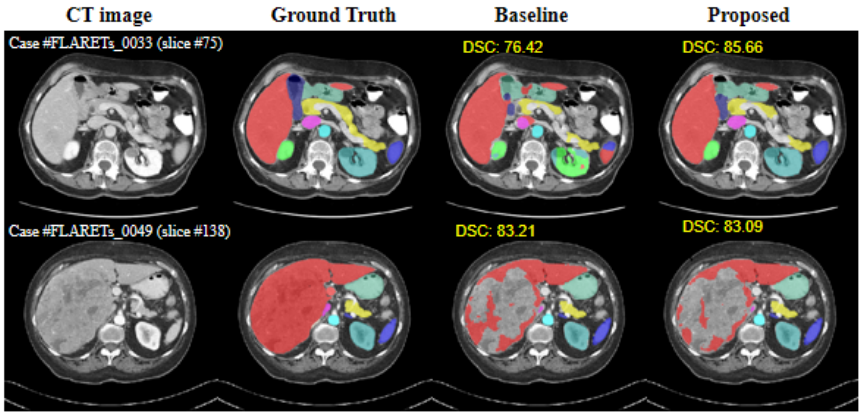

Fig. 6. Bad segmentation cases from 50 validation set.

Table 9 presents the time taken for several large CT scans, indicating that our method has a significant advantage in reducing inference time, especially for large CT scans.

Table 9. Overview of Ablation Experiment Results on inference speed. Proposed: Small nnU-Net + Z-RoI.

Dataset	Big nnU-Net(s)	Small nnU-Net(s)	Small nnU-Net + Z-RoI(s)
FLARETs_0001_0000	62.79	29.29	15.72
FLARETs_0010_0000	62.48	29.56	17.05
FLARETs_0050_0000	114.84	55.58	25.39

4.3 Segmentation efficiency results on validation set

We quantitatively evaluate the segmentation efficiency of our model based on running time, as shown in Table 10.

Table 10. Quantitative evaluation of segmentation efficiency in terms of the running time. Evaluation CPU: Intel(R) Core(TM) i5-12400F CPU @ 2.50GHz × 12.

Case ID	Image Size	Running Time (s)
0059	(512, 512, 55)	10.66
0005	(512, 512, 124)	6.62
0159	(512, 512, 152)	11.20
0176	(512, 512, 218)	9.68
0112	(512, 512, 299)	10.36
0135	(512, 512, 316)	10.53
0150	(512, 512, 457)	7.94
0134	(512, 512, 597)	12.77

4.4 Results on final testing set

4.5 Limitation and future work

In this study, the segmentation performance for small organs remains unsatisfactory, especially the boundary segmentation of these organs is not very clear. In future work, we will focus on improving the segmentation of these small organs and improve the boundary segmentation effect of these small organs, such as the gallbladder and adrenal glands. Additionally, we only determined the roi of the z-axis and did not consider the roi positioning of the x-axis and y-axis. In future work, we will explore the positioning scheme of the x-axis and y-axis to achieve the positioning scheme of the three axes.

5 Conclusion

To facilitate efficient semantic segmentation inference of abdominal CT scans on a CPU, this paper introduces a multi-scale knowledge distillation method to improve the performance of small models and design a weighted Composite loss function to alleviate class imbalance problem and Overcome the quality issues of pseudo-labels themselves. At inference, we further introduce a liver-based z-axis ROI localization strategy to accelerate inference. Quantitative and qualitative results demonstrate that our method can efficiently and flexibly learn information about multiple organs from the dataset. We validated our method on the MICCAI FLARE 2024 challenge dataset, proving that the proposed approach performs excellently in segmenting 13 different organs on a CPU.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2024 challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention. We thank all data owners for making the CT scans publicly available and CodaBench [29] for hosting the challenge platform.

Disclosure of Interests

The authors declare no competing interests.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.W., Georgescu, B., i Nieto, X.G., Gruen, F., Han, X., Heng, P.A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) 9
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2017) 2
3. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 9
4. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging **30**(9), 1323–1341 (2012) 9
5. Gatidis, S., Früh, M., Fabritius, M., Gu, S., Nikolaou, K., La Fougère, C., Ye, J., He, J., Peng, Y., Bi, L., et al.: The autopet challenge: Towards fully automated lesion segmentation in oncologic pet/ct imaging. Nature Machine Intelligence (in presss) (2024) 9

6. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022) 9
7. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 9
8. Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging. American Society of Clinical Oncology **38**(6), 626–626 (2020) 9
9. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. ArXiv Preprint **1503.02531** (2015) 3
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv Preprint **1704.04861** (2017) 2
11. Huang, Z., Wang, H., Ye, J., Niu, J., Tu, C., Yang, Y., Du, S., Deng, Z., Gu, L.: Revisiting nnu-net for iterative pseudo labeling and efficient sliding window inference. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Learning (2022) 6
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021) 2
13. 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
14. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022) 9
15. Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., Zhang, F., Liu, W., Pan, Y., Huang, S., Wang, J., Sun, M., Xu, W., Jia, D., Choi, J., Alves, N., de Wilde, B., Koehler, G., Wu, Y., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., Consortium, F.C., Wang, B.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. ArXiv Preprint **2308.05862** (2023) 2
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 9
17. Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B.: Segment anything in medical images and videos: Benchmark and deployment. arXiv preprint arXiv:2408.03322 (2024) 9
18. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Štern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.C., Qayyum, A., Conze,

P.H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022) 9

19. Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., Zhang, F., Liu, W., Pan, Y., Huang, S., Wang, J., Sun, M., Xu, W., Jia, D., Choi, J.W., Alves, N., de Wilde, B., Koehler, G., Wu, Y., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., the FLARE Challenge Consortium, Wang, B.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. Lancet Digital Health (2024) 9
20. Ma, J., Zhang, Y., Gu, S., Ge, C., Wang, E., Zhou, Q., Huang, Z., Lyu, P., He, J., Wang, B.: Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. arXiv preprint arXiv:2408.12534 (2024) 9
21. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022) 9
22. Oktay, O., Schlemper, J.S., Gopinath, A., Moeskops, P., et al.: Attention u-net: learning where to look for the pancreas. ArXiv Preprint **1804.03999** (2018) 2
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28 2
24. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 9
25. Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics **75**, 24–33 (2019) 6
26. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019) 2
27. Van Ginneken, B., Schaefer-Prokop, C.M., Prokop, M.: Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology **261**(3), 719–732 (2011) 2
28. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 9
29. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 16
30. Yushkevich, P.A., Gao, Y., Gerig, G.: Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3342–3345 (2016) 9

31. Zhou, P., Mai, L., Zhang, J., Xu, N., Wu, Z., Davis, L.S.: M2kd: Multi-model and multi-level knowledge distillation for incremental learning. ArXiv Preprint **2003.09112** (2020) 3

Table 11. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6) 6	Number
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	6
Pre-processing	5
Strategies to improve model inference	5-8
Post-processing	8
The dataset and evaluation metric section are presented	9
Environment setting table is provided	1
Training protocol table is provided	2
Ablation study	7-9
Efficiency evaluation results are provided	3
Visualized segmentation example is provided	5 6
Limitation and future work are presented	Yes
Reference format is consistent.	Yes