Lightweight nnU-Net with Knowledge Distillation and Multi-Threaded Optimization for Abdominal Organ Segmentation

Kai Hou[†][0009−0005−9152−4436], Yue Liu[†][0000−0002−4032−6318]and Yinyin Luo[0009−0009−0921−6047], Bingquan Huang[0009−0002−8195−3358], Bin Tang[0009−0006−5833−8604], and Gang Fang[*][0000−0001−9847−114X]

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 models have achieved remarkable success in medical image segmentation, their deployment on resource-constrained environments remains challenging due to substantial computational and memory requirements, particularly for 3D medical images. Existing lightweight models often sacrifice segmentation accuracy significantly to reduce computational overhead.

To address this challenge, we propose a comprehensive lightweight optimization framework based on nnU-Net that maintains high segmentation accuracy while dramatically reducing computational requirements. Our main contributions include: (1) a lightweight network architecture that replaces standard 3D convolutions with depthwise separable convolutions and incorporates bottleneck residual blocks, reducing model parameters to 157.59K while preserving the feature representation capability; (2) pyramid pooling modules for enhanced multi-scale feature extraction and improved boundary precision; (3) a knowledge distillation strategy where a teacher network (original nnU-Net) transfers knowledge to our lightweight student network through feature-level and output-level distillation losses; and (4) a multi-threaded inference optimization system that parallelizes post-processing operations using 12 threads, achieving 2-4× speedup in post-processing.

Comprehensive experiments on MICCAI FLARE 2025 validation set validate the effectiveness of our approach. The proposed method achieves an average organ Dice Similarity Coefficient (DSC) of 90.48% and Normalized Surface Dice (NSD) of 96.56%, and the validation score improved by 0.9 points. Our method ranked 1st in the MICCAI FLARE 2025 Task 2 online validation leaderboard. The code is available at: https://github.com/houkainiubi/flare25task2_hk.git

Keywords: Knowledge Distillation · Depthwise Separable Convolution · Pyramid Pooling · Multi-threading Optimization · Bottleneck Residual Blocks.

1 Introduction

Medical image segmentation plays a crucial role in computer-aided diagnosis, treatment planning, and clinical decision-making [14]. Accurate segmentation of anatomical structures and pathological regions from medical images enables quantitative analysis, disease progression monitoring, and personalized medicine. However, deploying robust segmentation models in real-world clinical environments presents significant computational and practical challenges that must be addressed to ensure widespread clinical adoption [11].

The computational constraints in clinical settings pose one of the greatest challenges for medical image segmentation systems. Unlike research [7] environments with access to high-performance GPUs and abundant computational resources [12], clinical deployment scenarios often operate under strict hardware limitations. In this challenge, models must perform inference exclusively on CPU architectures, which typically offer significantly lower computational throughput compared to GPU-accelerated systems. This constraint dramatically impacts the feasibility of deploying computationally intensive deep learning models that have become standard in medical image analysis research.

Memory limitations [27] further compound the deployment challenges, as the available runtime memory is restricted to merely 8GB during inference. This constraint is particularly challenging for medical image segmentation tasks, which often involve processing high-resolution volumetric data or large-scale 2D images. Traditional segmentation networks, especially those based on U-Net architectures and their variants, typically require substantial memory for storing intermediate feature maps, particularly in the encoder-decoder pathways. The memory bottleneck becomes even more severe when processing multiple cases simultaneously or when dealing with 3D volumetric segmentation tasks that inherently demand more memory resources.

Perhaps the most significant architectural constraint imposed by this challenge is the prohibition of two-stage cascade models, specifically those employing localization-then-segmentation paradigms. Many state-of-the-art medical segmentation systems rely on cascade approaches where an initial network localizes the region [2] of interest, followed by a specialized segmentation network that focuses on the identified region. This two-stage strategy has proven highly effective in improving both accuracy and computational efficiency by reducing the search space for the segmentation network. However, the necessitates end-to-end single-stage models that achieve comparable performance under strict computational constraints.

To meet the requirements of Task 2 in the MICCAI FLARE 2025 Challenge, which involves deploying advanced 3D abdominal CT segmentation models in non-GPU environments while maintaining high accuracy, this paper introduces improvements to the nnU-Net model, along with knowledge distillation and post-processing strategies. Specifically, Task 2 of the MICCAI FLARE 2025 Challenge requires semantic segmentation of abdominal organ CT images using CPU-based algorithms on a laptop with an 8GB memory limit. This task utilizes the same dataset as MICCAI FLARE 2024, involving the segmentation of 13 organs from

CT images provided by over 20 medical groups, with organ labels as shown in Figure 1 . The dataset includes 2050 cases for model training, 250 cases for val-

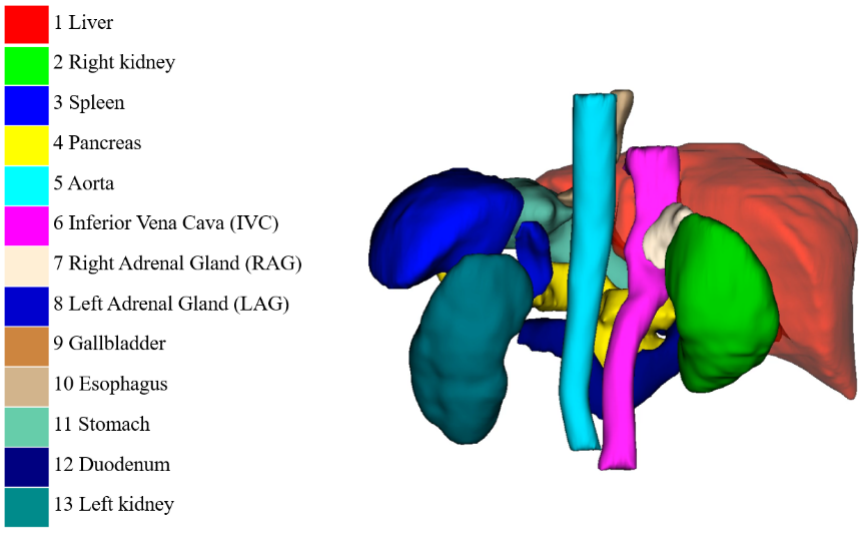

Fig. 1. Semantic labels of the 13 abdominal organs in FLARE 2025 Task2.

idation, and 300 test cases for final evaluation. The evaluation metrics include the Dice Similarity Coefficient (DSC), Normalized Surface Dice (NSD), and the number of seconds required for inferring a single CT image (runtime).
Medical image segmentation has witnessed remarkable progress with the advent of deep learning, particularly with the introduction of U-Net and its variants. The nnU-Net framework has emerged as a gold standard for medical image segmentation, achieving state-of-the-art performance across multiple medical imaging tasks through its self-configuring pipeline and robust architecture design. However, the computational demands of such high-performance models pose significant challenges for deployment in resource-constrained clinical environments.

Recent efforts in model compression and efficiency optimization have explored various approaches including network pruning, quantization, and knowledge distillation. Knowledge distillation, originally proposed by Hinton et al., has proven particularly effective in transferring knowledge from large teacher networks to compact student networks while preserving performance. In the context of medical image segmentation, several works have demonstrated the effectiveness of distillation-based approaches for creating efficient models suitable for clinical deployment.

Lightweight network architectures, such as MobileNets and EfficientNets, have introduced depthwise separable convolutions and inverted residual blocks to significantly reduce computational complexity while maintaining representa-

tional capacity [10]. However, the direct application of these architectures to 3D medical image segmentation remains challenging due to the unique characteristics of volumetric medical data and the need for precise boundary delineation [14].

Following prior works that leverage lightweight architectures for efficient inference [25], we propose a comprehensive optimization framework that integrates multiple architectural innovations with advanced knowledge distillation strategies. Our approach is specifically designed to address the stringent computational constraints of the MICCAI FLARE 2025 Challenge Task 2 while maintaining high segmentation accuracy.

Lightweight Network Architecture with Advanced Components: We introduce a novel lightweight network architecture that replaces standard 3D convolutions with depthwise separable convolutions and incorporates bottleneck residual blocks to dramatically reduce model parameters. The architecture is further enhanced with pyramid pooling modules [26] as shown in Figure 2 to capture

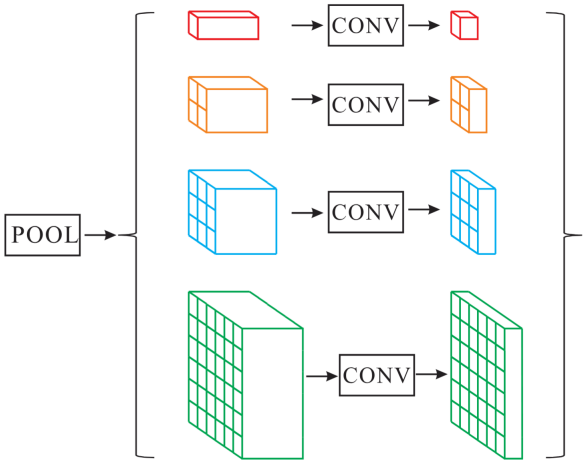

Fig. 2. Pyramid Pooling Block.

multi-scale contextual information, enabling improved boundary precision for organs of varying sizes. This multi-component design addresses the fundamental trade-off between model efficiency and representational capacity in 3D medical image segmentation.

Multi-Scale Knowledge Distillation Strategy: To preserve the segmentation performance of the lightweight student model, we employ a sophisticated knowledge distillation framework where a teacher network (original nnU-Net) transfers its knowledge to our lightweight student network through both feature-level and output-level distillation losses. Specifically, we implement a temperature-scaled knowledge distillation loss that combines KL divergence, MSE, and attention

transfer mechanisms to ensure comprehensive knowledge transfer across different scales and representation levels. This multi-faceted distillation approach enables the student model to retain the teacher's performance while operating with significantly reduced computational resources.

System-Level Inference Optimization: Beyond architectural improvements, we implement a comprehensive multi-threaded inference optimization system that parallelizes the post-processing pipeline. Our optimization strategy includes parallel resampling operations, segmentation post-processing, and file I/O operations using up to 12 threads, achieving 2-4× speedup in the post-processing stage. The system incorporates dynamic resource management, memory optimization strategies, and robust error handling to ensure stable performance across diverse computational environments and varying input sizes.

Robust Training and Loss Design: To improve the segmentation performance of small organs and address potential issues with training stability, we design a weighted composite loss function that combines Dice loss and cross-entropy loss with careful weighting strategies. The loss function is specifically optimized for the knowledge distillation framework to ensure stable convergence and effective knowledge transfer.

To summarize, our main contributions are: (1) A comprehensive lightweight architecture combining depthwise separable convolutions, bottleneck Residual Block Using Deep Separable Convolution, and pyramid pooling modules specifically optimized for 3D medical image segmentation; (2) A multi-scale knowledge distillation strategy that effectively transfers knowledge from teacher to student networks through multiple distillation mechanisms; (3) A system-level optimization framework featuring multi-threaded post-processing that achieves significant speedup while maintaining robustness; (4) Comprehensive experimental validation demonstrating that our approach successfully balances segmentation accuracy and computational efficiency, making advanced 3D medical image segmentation feasible for deployment in resource-limited clinical environments.

2 Method

2.1 Preprocessing

- **Resample:** We resample the pixel spacing to (1.9735,1.5380, 1.5380) for all cases.

2.2 Proposed Method

Model Architecture and Improvements In this work, we propose an improved nnUNet-based segmentation framework that integrates depthwise separable convolutions, bottleneck residual blocks, pyramid pooling, and a knowledge distillation mechanism. The overall network follows the encoder–decoder paradigm with skip connections between symmetric layers, as illustrated in Figure 3 .

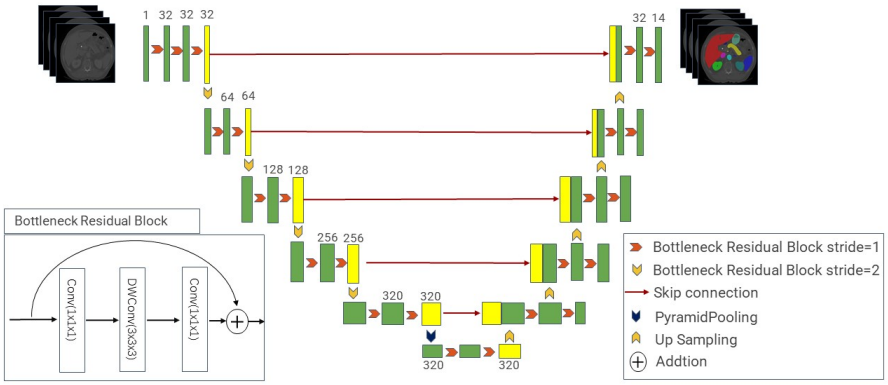

Fig. 3. Teacher model of the proposed method.

The encoder is composed of multiple stages, each containing a stack of Bottleneck Residual Blocks. Each block employs a 1×1 convolution for channel reduction, followed by a depthwise separable 3×3 convolution to extract spatial features, and another 1×1 convolution for channel restoration. A residual connection is added between the input and output of the block to facilitate gradient propagation and mitigate degradation in deeper layers. Two types of bottleneck residual blocks are utilized: one with stride = 1 for feature refinement and one with stride = 2 for spatial downsampling.

In the bottleneck stage of the encoder, a Pyramid Pooling Module (PPM) is incorporated to capture multi-scale contextual information. The PPM aggregates features from different spatial scales and fuses them back into the main feature stream, enhancing the model's ability to segment organs of varying sizes.

The decoder mirrors the encoder structure and progressively upsamples the feature maps to the original resolution using bilinear upsampling layers, followed by bottleneck residual blocks for refinement. Skip connections between encoder and decoder stages preserve fine-grained details lost during downsampling.

To enhance computational efficiency while maintaining accuracy, all standard convolutions in the network are replaced with depthwise separable convolutions, significantly reducing the number of parameters and floating-point operations.

Knowledge Distillation Strategy: A Teacher–Student learning scheme is adopted to further boost segmentation performance. First, a large-capacity Teacher model, initialized with 32 base channels, is trained using a compound loss combining Dice loss and Cross-Entropy loss. The Teacher network achieves high segmentation accuracy on the FLARE25 dataset with pseudo-labels.

Subsequently, a lightweight Student model is constructed, initialized with 16 base channels, as illustrated in Figure 4 . To ensure feature alignment for knowl-

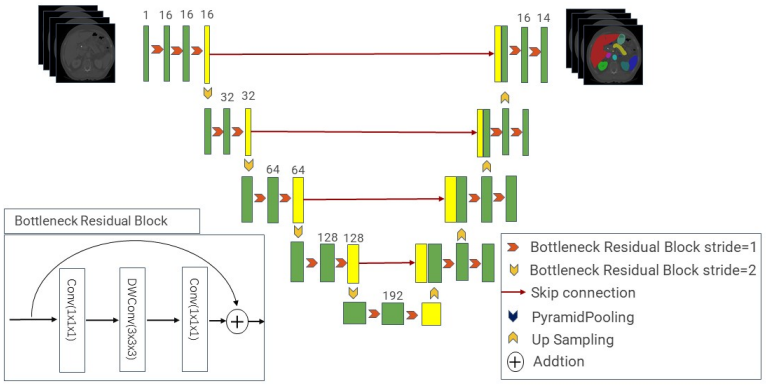

Fig. 4. Student model of the proposed method.

edge transfer, an additional non-resolution-changing layer is inserted in the Student encoder to match the depth of the Teacher model. Multi-scale feature distillation is performed at the two lowest encoder layers using Mean Squared Error (MSE) loss. The distillation losses from the two scales are summed, weighted by w=0.5, and added to the Student's supervised loss to form the final optimization objective.

Both Teacher and Student networks adopt deep supervision, where auxiliary segmentation outputs are generated at multiple decoder stages to stabilize training. The pseudo-labels used for both models are generated from 2,000 unlabeled cases based on the winning solution of FLARE 2025, without further post-processing.

This combination of architectural optimization and multi-scale knowledge distillation allows the Student model to achieve competitive accuracy while significantly reducing computational cost, making it more suitable for deployment in resource-constrained environments.

2.3 Post-processing

Multi-threaded Post-processing Pipeline To address the computational bottleneck in the post-processing stage, we implemented a multi-threaded optimization pipeline that significantly accelerates the resampling and segmentation export processes.

Parallel Resampling Strategy The most time-consuming operation in post-processing is resampling predictions from low resolution back to original resolution. We implemented a class-wise parallel resampling approach: Thread alloca-

tion: Dynamically allocates up to 12 threads based on physical CPU cores and memory constraints Class-wise parallelization: Each segmentation class is processed independently using separate threads Memory-aware processing: Thread count is automatically reduced for large datasets (>8GB) to prevent memory overflow Error handling: Robust fallback mechanisms ensure processing continues even if individual threads fail.

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 [21], KiTS [8,9], autoPET [6,5], AMOS [13], AbdomenCT-1K [20], TotalSegmentator [22], and past FLARE challenges [17,18,19]. 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 [24], nnU-Net[11], MedSAM [15,16], and Slicer Plugins [4,16]. In additional to use all training cases for model development, we also added a coreset track where participants can select 50 cases from the training set for model development in an automatic way.

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 20.04 LTS or Windows 11
CPU	Intel(R) Core(TM) i9-13900HX CPU@5.40GHz
RAM	8×2GB; 2933MT/s
Programming language	Python 3.12
Deep learning framework	torch 2.0, torchvision 0.2.2
Specific dependencies	nnU-Net
Code	https://github.com/houkainiubi/flare25task2_hk.git

Training protocols Please describe at least the following aspects:

1. Data augmentation (Based on the winning solutions in FLARE 2021 [17], we recommend using extensive data augmentation)
2. patch sampling strategy
3. optimal model selection criteria

Table 2. Training protocols.

Network initialization	
Batch size	2
Patch size	80×160×160
Total epochs	2700
Optimizer	AdamW with weight decay(μ = 1e -5)
Initial learning rate (lr)	0.01
Lr decay schedule	halved by 200 epochs
Training time	45 hours
Loss function	DiceLoss and CELoss
Number of model parameters	157.59K[1]
Number of flops	60G[2]
CO_2eq	0.5 Kg[3]

4 Results and discussion

4.1 Quantitative results on validation set

The quantitative evaluation results are presented in Table 3 , which indicate that the proposed method yields highly promising outcomes in segmenting major organs such as the liver, spleen, kidneys, and stomach. Nevertheless, segmenting smaller organs remains a significant challenge that demands further attention, especially for extremely small organs with indistinct boundaries, such as the adrenal glands and duodenum.

We conducted ablation experiments on the constructed base model, the model with one-time knowledge distillation, and the model with two-time knowledge distillation. The results show that the base model without knowledge distillation achieves the highest accuracy but has the slowest inference speed. The model with two-time knowledge distillation yields the lowest accuracy but boasts the fastest inference speed. Meanwhile, we conducted ablation experiments on both the public validation set and the online validation set. The results are shown in Table 4 and Table 5.

Table 3. Quantitative evaluation results.

Target	Public Validation		Online Validation		Testing	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	97.87	92.79	97.90	99.17		
Right Kidney	95.20	95.74	95.19	96.44		
Spleen	96.24	96.50	95.92	97.77		
Pancreas	88.90	96.77	89.73	97.82		
Aorta	96.79	98.34	94.90	98.39		
Inferior vena cava	89.30	89.24	90.98	93.71		
Right adrenal gland	79.99	90.94	82.04	95.34		
Left adrenal gland	75.45	85.25	81.43	94.20		
Gallbladder	79.90	79.70	84.20	85.98		
Esophagus	87.22	95.38	86.16	95.68		
Stomach	92.37	94.92	94.64	97.79		
Duodenum	80.10	92.44	85.60	96.50		
Left kidney	91.95	93.08	95.09	96.33		
Average	88.56	92.79	90.29	95.78		

Table 4. Overview of Ablation Experiment Results in Public Validation. Proposed:Base+Distill+Distill2nd

Target	Base		Base+Distill		Base+Distill2nd	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	98.15	98.47	97.87	92.79	97.39	97.30
Right Kidney	95.46	96.15	95.20	95.74	90.89	90.35
Spleen	97.16	97.25	96.24	96.50	95.55	95.41
Pancreas	89.77	97.21	88.90	96.77	79.39	89.78
Aorta	97.09	98.64	96.79	98.34	93.02	94.66
Inferior vena cava	90.58	90.45	89.30	89.24	80.90	77.55
Right adrenal gland	83.99	93.15	79.99	90.94	55.17	65.31
Left adrenal gland	81.25	89.71	75.45	85.25	47.15	52.72
Gallbladder	79.43	80.16	79.90	79.70	78.06	75.44
Esophagus	89.06	96.17	87.22	95.38	73.58	83.89
Stomach	92.71	95.00	92.37	94.92	88.38	91.51
Duodenum	81.48	92.67	80.10	92.44	65.36	80.85
Left kidney	92.84	93.66	91.95	93.08	90.10	90.18
Average	89.92	93.75	88.56	92.79	79.61	83.47

Table 5. Overview of Ablation Experiment Results in Online Validation. Proposed:Base+Distill+Distill2nd

Target	Base		Base+Distill		Base+Distill2nd	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	98.04	99.40	97.90	99.17	97.27	98.44
Right Kidney	95.87	91.11	95.19	96.44	94.88	96.07
Spleen	96.00	97.81	95.92	97.77	94.91	96.50
Pancreas	90.19	97.95	89.73	97.82	81.27	92.09
Aorta	94.87	98.29	94.90	98.39	91.97	94.06
Inferior vena cava	91.44	94.30	90.98	93.71	80.92	79.23
Right adrenal gland	83.89	96.35	82.04	95.34	61.01	71.63
Left adrenal gland	84.87	96.63	81.43	94.20	54.61	64.04
Gallbladder	83.52	85.78	84.20	85.98	78.98	77.31
Esophagus	87.87	96.80	86.16	95.68	73.58	84.81
Stomach	94.89	97.94	94.64	97.79	91.68	95.04
Duodenum	86.31	96.70	85.60	96.50	69.02	86.29
Left kidney	95.27	96.51	95.09	96.33	94.60	96.05
Average	91.00	96.27	90.29	95.78	81.90	87.04

4.2 Qualitative results on validation set

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

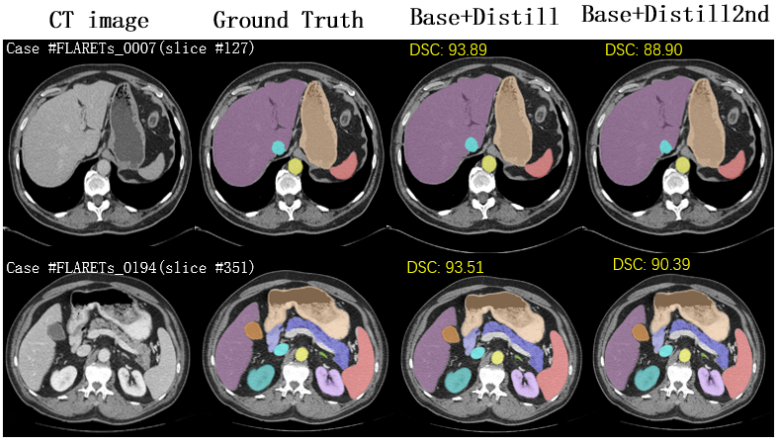

Fig. 5. Good segmentation cases from 200 validation set.

Good segmentation cases: Figure 5 illustrates representative segmentation results for two cases. In FLARETs_0007, the Base+Distill method achieves

a close alignment with the ground truth across all major abdominal organs, with a high DSC of 93.89. Conversely, the Base+Distill2nd method shows lower segmentation accuracy, where the liver and stomach boundaries become less precise, leading to a notable DSC drop to 88.90. In FLARETs_0194, both methods generally produce satisfactory segmentation results, yet Base+Distill again demonstrates superior accuracy (DSC: 93.51). In contrast, Base+Distill2nd fails to capture finer organ structures, particularly at the pancreas and adjacent vascular regions, contributing to its reduced DSC of 90.39. These findings suggest that while both methods are capable of segmenting large organs effectively, Base+Distill maintains better sensitivity to organ boundaries and small structures, whereas Base+Distill2nd tends to lose detail, resulting in degraded segmentation quality.

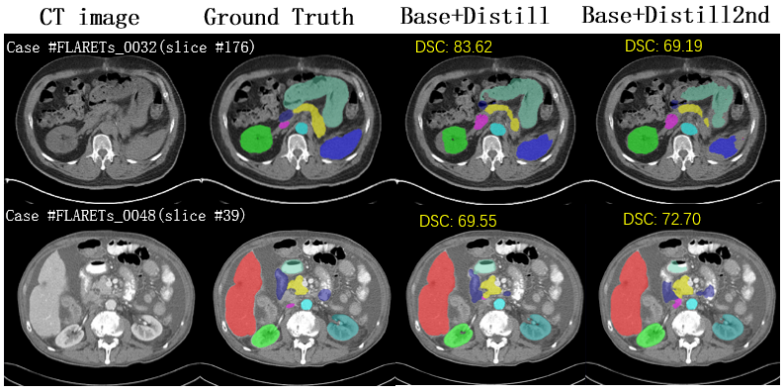

Fig. 6. Bad segmentation cases from 50 validation set.

Bad segmentation cases: Figure 6 presents two examples where the segmentation performance degrades significantly. In FLARETs_0032, the Base+Distill method maintains relatively good delineation of the pancreas and surrounding organs with a DSC of 83.62, whereas Base+Distill2nd shows noticeable errors, particularly at the pancreas and duodenum boundaries, resulting in a substantial DSC drop to 69.19. In FLARETs_0048, both methods struggle with accurate segmentation. The Base+Distill method exhibits under-segmentation of the spleen and kidneys, leading to a DSC of 69.55, while Base+Distill2nd slightly improves the overall score (72.70) but still produces misclassifications in vascular regions and small organ structures. These cases highlight that both approaches face difficulties in segmenting organs with complex shapes or low contrast against adjacent tissues, with Base+Distill2nd being more prone to boundary inaccuracies, particularly for smaller structures.

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 6 and Table 7. It can be clearly seen that the model reasoning speed after a knowledge distillation is greatly accelerated. Even if the model accuracy after a knowledge distillation is slightly lost compared to the basic model, it is worth it.

Table 6. Segmentation time on the Base+Distill model. Quantitative evaluation of segmentation efficiency in terms of the running time. Evaluation CPU: Intel Xeon(R) W-2133 CPU @ 3.60GHz × 12.

Case ID	Image Size	Running Time (s)
0007	(512, 512, 215)	116.71
0027	(512, 512, 169)	61.29
0029	(512, 512, 171)	63.08
0036	(512, 512, 91)	31.65
0058	(512, 512, 56)	52.75
0063	(512, 512, 361)	113.41
0071	(512, 512, 108)	60.21
0164	(512, 512, 114)	161.21
0189	(512, 512, 89)	50.17
0190	(512, 512, 101)	117.29

Table 7. Segmentation time on the Base model.Quantitative evaluation of segmentation efficiency in terms of the running time. Evaluation CPU: Intel Xeon(R) W-2133 CPU @ 3.60GHz × 12.

Case ID	Image Size	Running Time (s)
0007	(512, 512, 215)	146.79
0027	(512, 512, 169)	123.66
0029	(512, 512, 171)	122.64
0036	(512, 512, 91)	70.87
0058	(512, 512, 56)	85.37
0063	(512, 512, 361)	175.61
0071	(512, 512, 108)	81.05
0164	(512, 512, 114)	202.47
0189	(512, 512, 89)	97.47
0190	(512, 512, 101)	174.05

4.4 Results on final testing set

4.5 Limitation and future work

Although the proposed modifications to the nnU-Net framework, including the integration of depthwise separable convolutions, pyramid pooling, bottleneck residual blocks, and multi-threaded post-processing, have demonstrated significant improvements in both segmentation accuracy and inference efficiency, several limitations remain. First, the current approach still relies on the general preprocessing pipeline of nnU-Net, which may not optimally capture the region of interest (ROI) in highly variable abdominal CT scans. This limitation could lead to redundant computations and reduced robustness in cases where organ boundaries are subtle or imaging quality is degraded. Second, while the proposed architectural changes improved overall performance, the model occasionally struggles with small or low-contrast structures, suggesting that further refinement in feature extraction and multi-scale representation is necessary. Finally, the evaluation was constrained by the competition dataset, and generalization to diverse clinical datasets remains to be validated.

In future work, we aim to address these issues by developing a more adaptive preprocessing strategy that incorporates precise ROI localization prior to segmentation. By narrowing the target area during preprocessing, the model can potentially reduce computational overhead and focus more effectively on relevant anatomical structures. Additionally, integrating attention-based mechanisms and exploring self-supervised pretraining on large-scale medical imaging datasets could further enhance the model's ability to capture fine-grained details and generalize across different imaging modalities and institutions.

5 Conclusion

In this study, we presented an enhanced segmentation framework based on nnU-Net, designed and optimized for the medical image segmentation challenge. Through architectural modifications—namely the introduction of depthwise separable convolutions, pyramid pooling modules, and bottleneck residual connections—combined with an efficient multi-threaded post-processing scheme, our method achieved substantial gains in both segmentation accuracy and computational efficiency. The experimental results demonstrated the effectiveness of these modifications in handling complex abdominal CT scans and improving organ delineation compared to the baseline.

Despite its limitations, the proposed framework offers a practical balance between accuracy and efficiency, making it well-suited for large-scale medical imaging tasks. Our findings further underscore the flexibility of nnU-Net as a foundation for innovation in medical image analysis. Future work will focus on advancing preprocessing strategies, particularly precise ROI localization, to further reduce computational burden and improve segmentation robustness in challenging scenarios. Overall, this work highlights the potential of tailored architectural and system-level optimizations to push the performance boundaries of automated medical image segmentation.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2025 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 [23] 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) 8
2. Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., et al.: Automatic liver and lesion segmentation using cascaded fully convolutional neural networks. In: MICCAI. pp. 415–423. Springer (2016) 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) 8
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) 8
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 press) (2024) 8

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) 8
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 2
8. 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) 8
9. 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) 8
10. Howard, A.G., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. In: arXiv preprint arXiv:1704.04861 (2017) 4
11. 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, 8
12. Isensee, F., et al.: nnu-net: Breaking the spell on successful medical image segmentation. In: arXiv preprint arXiv:1904.08128 (2019) 2
13. 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) 8
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: Deep learning in medical image analysis: A survey. Medical image analysis **42**, 60–88 (2017) 2, 4
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 8
16. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025) 8
17. 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) 8, 9
18. 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) 8

19. 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) 8
20. 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) 8
21. 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) 8
22. 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) 8
23. 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) 15
24. 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) 8
25. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. CVPR pp. 6848–6856 (2018) 4
26. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017) 4
27. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., Zhang, J.: Deep learning for edge computing: A review. Proceedings of the IEEE **107**(8), 1655–1674 (2019) 2

Table 8. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	6
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	6-7
Pre-processing	5
Strategies to improve model inference	5-6-7
Post-processing	7
The dataset and evaluation metric section are presented	8
Environment setting table is provided	8
Training protocol table is provided	9
Ablation study	10-11
Efficiency evaluation results are provided	13
Visualized segmentation example is provided	11-12
Limitation and future work are presented	13-14
Reference format is consistent.	Yes