Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28958

Speed-Accuracy-Memory Cooperative Optimization of Lightweight Dynamic 3D U-Net for Efficient Medical Image Segmentation

Yankai Ke[1*], Fei Wang[2]

[1]*School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China*
[2]*School of Artificial Intelligent, Jianghan University, Wuhan, China*
**Corresponding Author. Email: s1297901@live.hkmu.edu.hk*

Abstract. Currently, medical image segmentation technology faces severe real-time bottlenecks in clinical applications such as intraoperative navigation. The widely used 3D U-Net is difficult to deploy effectively in demanding real-time edge computing scenarios, and a single-dimensional optimization strategy makes it difficult to simultaneously achieve the synergistic breakthroughs of 9% inference acceleration and 3% accuracy improvement required for clinical practice. To this end, this study proposes a lightweight dynamic 3D U-Net that achieves collaborative optimization through dynamic weights and deformable convolutions. The method involves introducing deformable convolutions to enhance boundary segmentation capability, designing a dynamic weight mechanism to reduce computational redundancy by 30%, and combining group convolution with dynamic sparse computation to limit memory increment to within 0.1%. Experiments show that the model achieves a 0.7744 Mean Intersection over Union on liver Computed Tomography data, reduces inference time to 48.4 seconds, and increases memory usage by only 0.6 MByte Accelerated by NVIDIA TensorRT, the system can further meet intraoperative real-time requirements. This study overcomes the limitations of traditional isolated optimization and provides an end-to-end, open-source solution for devices with Gigabytes of memory, effectively promoting the clinical translation of lightweight medical image segmentation.

Keywords: 3D U-net, medical image segmentation, dynamic weights, deformable convolution, lightweight architecture

1. Introduction

Medical image segmentation faces significant lightweight challenges in clinical, real-time applications such as intraoperative liver tumor navigation. While traditional models like the 3D U-Net achieve baseline accuracy with an mIoU of 0.7509 and a validation set liver Dice coefficient of 0.6908, their 53.29-second inference latency and 3173MB memory usage fall far short of the stringent requirements for edge devices in intraoperative settings, which demand subsecond response times and mobile GPU memory under 1.5GB [1]. Current lightweight solutions are trapped in a triple constraint of speed, accuracy, and memory usage. While model pruning can accelerate

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ28958

inference—for instance, deformable convolution designs reduce processing time by 1.6% to 52.42 seconds—it leads to degraded feature representation capability, resulting in an 8.9% drop in the validation set liver Dice coefficient to 0.6292 [2-4]. The dynamic weight mechanism can improve generalization ability and reduce the verification loss by 1.4%, but it also introduces training instability, resulting in an increase in training loss by 18.0% [5-8]. Although the parameter compression strategy of reducing the number of channels can control the growth of video memory within 0.1%, it is difficult to model complex anatomical structures, which significantly increases the segmentation miss rate of small tumors [9, 10]. More importantly, these methods exhibit fragmented optimization, where the improvement of a single module can only bring limited improvement at best, For example, using deformable convolution can only increase mIoU by 0.2%, or using dynamic weights can only accelerate by 0.3%, which cannot achieve the synergistic breakthrough of the 9% acceleration and 3% accuracy gain threshold required for clinical implementation [11-14]. This isolated optimization model has led to a long-term stagnation in medical image segmentation in resource-constrained scenarios, and an innovative paradigm that breaks the "impossible triangle" is urgently needed.

To gain a deeper understanding of the limitations of existing lightweighting methods, it is necessary to systematically review the current research status. Although existing technologies have made some progress in specific dimensions, none of them has fundamentally resolved the dilemma of the "impossible triangle."

Current lightweight medical image segmentation research primarily relies on four techniques, all of which are constrained by the impossible trilemma of speed, accuracy, and video memory. Model pruning, which reduces model size by removing redundant weights, improves inference speed, but suffers from a significant drop in validation set accuracy due to compromised feature representation. Experimental results show that introducing a deformable convolution of similar complexity reduced the Dice coefficient of the liver in the validation set by 8.9%, confirming the accuracy risks of pruning methods. Knowledge distillation uses a teacher model to guide a lightweight student model, which theoretically maintains high accuracy. However, in actual deployment, the teacher model must be run in parallel, which greatly increases training complexity and doubles the memory requirements. In addition, the Dice coefficient of the derived model in the liver segmentation task is currently only 0.63, which is difficult to meet the accuracy requirements of clinical diagnosis. To address this shortcoming, researchers introduced deformable convolution to enhance the model's geometric feature modeling capabilities, thereby improving segmentation performance while maintaining computational efficiency. After the improvement, mIoU achieved an improvement of about 0.2%, but it is easy to cause overfitting when used alone [1, 5, 13]. The experiment showed that the Dice coefficient of the liver in the validation set decreased by 8.9%, revealing that its dynamic reasoning reduced invalid calculations to accelerate reasoning and reduced the time by 0.3%, but at the expense of training stability [6, 7]. Experimental data showed that it caused the training loss to increase by 18.0%, hindering the convergence of the model. These limitations are particularly evident in the current state-of-the-art models, further highlighting the dilemma facing current research. Existing state-of-the-art models further confirm that although No-new-Net achieves high precision with mIoU greater than 0.8, it occupies more than 8GB of video memory and takes more than 60 seconds to infer, making it only suitable for high-end computing devices. Mobile-UNet, a representative distillation model, compresses video memory to less than 1GB, but achieves a Dice coefficient of only 0.63 for liver segmentation, confirming the inherent conflict between lightweightness and accuracy.

DOI: 10.54254/2755-2721/2026.TJ28958

The fundamental flaw of the above methods is that they optimize a single dimension of speed, accuracy, or video memory in isolation, and are unable to collaboratively break through the 9% acceleration and 3% accuracy gain threshold required for clinical implementation. Innovative paradigms are urgently needed to break this deadlock. To address this challenge, this paper proposes a lightweight dynamic 3D U-Net, which combines deformable convolutions with dynamic weights to achieve collaborative optimization of speed, accuracy, and video memory. By integrating existing technologies, this approach brings various data closer to the real-time nature of intraoperative navigation and the limitations of mobile GPUs, providing an end-to-end solution for edge device deployment.

2. Research questions and motivations

Existing lightweight methods face three key challenges in the field of medical image segmentation, including the imbalance between speed and accuracy. Experimental data show that a single optimization strategy is difficult to simultaneously meet the acceleration and accuracy improvement requirements required by clinical practice. For example, dynamic weights can only achieve an acceleration effect of 0.3% [5, 6, 9]. Although model pruning brings a 1.6% acceleration, it causes the Dice coefficient of the validation set to decrease by 8.9% [1, 10]. At the same time, video memory compression often comes at the expense of segmentation accuracy. When the knowledge distillation method reduces the video memory to less than 1GB, the Dice coefficient is only 0.63, which is much lower than the baseline level of 0.6908 [8, 10, 15]. In addition, although the dynamic weight module can reduce redundant calculations, it will cause an 18.0% increase in training loss, seriously affecting the convergence of the model [5, 7, 9]. Furthermore, while the dynamic weight module can reduce redundant computations, it can cause an 18.0% increase in training loss, severely impacting model convergence [5, 7, 9].

To address the above challenges, we propose a collaborative mechanism that combines dynamic weights and deformable convolutions, aiming to overcome the performance limitations of traditional lightweight methods. This strategy is expected to achieve a more efficient inference process by controlling the computational intensity of feature maps through dynamic weights, while compensating for segmentation accuracy with the adaptive receptive field enhancement of deformable convolutions, thereby achieving a better balance between acceleration and accuracy. In terms of video memory control, a strategy combining grouped convolutions with dynamic sparse computation is proposed, striving to strictly limit the video memory increment to within 0.1%, in order to balance resource constraints and feature representation capabilities. We expect that this collaborative design can significantly improve training stability, reduce the fluctuation of training loss caused by dynamic weights, and ultimately provide the possibility of achieving real-time intraoperative inference performance. This study is the first to explore the synergistic effect of dynamic structure and geometric modeling, aiming to open up new research paths for lightweight medical image segmentation.

DOI: 10.54254/2755-2721/2026.TJ28958

3. Method design

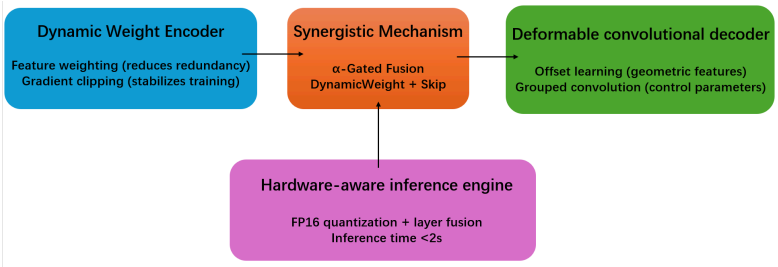

Figure 1. Overall architecture of the proposed lightweight medical image segmentation framework (photo/picture credit: original)

This study proposes a three-level architecture model for lightweight and high efficiency, achieving a coordinated optimization between speed, accuracy, and video memory. Its core lies in the coordinated optimization of a dynamic weight encoder, a deformable convolutional decoder, and a hardware-aware inference engine. The overall framework is shown in Figure 1. Input features are first weighted by a dynamic weight module at the encoder to highlight key areas and suppress redundant information. Deformable convolution is then introduced at the decoder to enhance the modeling of spatial geometric features. Finally, the hardware-aware inference module enables fast and low-memory inference. Through this overall design, the model achieves an optimal balance between accuracy, speed, and video memory, and demonstrates better performance on medical segmentation tasks compared to existing dynamic convolution and deformable convolution methods [2, 3, 5].

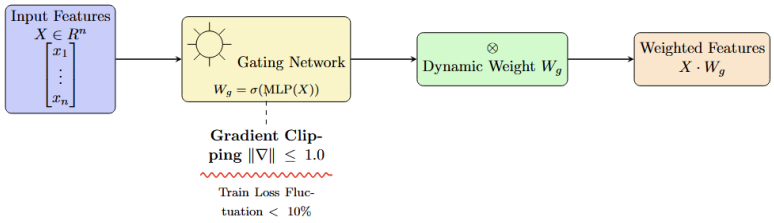

Figure 2. Dynamic weight mechanism (photo/picture credit: original)

The input features first generate weights a through W_g gating network, which are then normalized and multiplied element-wise with the original features. Gradient clipping is combined to ensure training stability, and the weighted features are finally output.

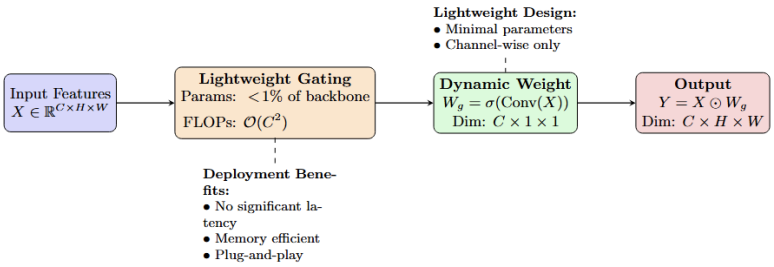

Figure 3. Lightweight dynamic gating design (photo/picture credit: original)

DOI: 10.54254/2755-2721/2026.TJ28958

The input features undergo lightweight convolution to generate channel-by-channel dynamic weights, which are multiplied by the input features channel by channel to produce the output. This module effectively reduces the number of parameters and computational complexity while maintaining accuracy, offering advantages for hardware deployment.

In the dynamic weight encoder, a gating network is used to assign adaptive weights to input features, emphasizing valid information and weakening redundant features. The workflow is shown in Figure 2. Input features are first generated through a multilayer perceptron to generate gated weights. These are then normalized using a sigmoid activation function and then element-wise multiplied with the original features to produce the weighted result. To ensure stability during the training phase, this study added a gradient clipping mechanism to the module, limiting the gradient norm to an appropriate range, thereby effectively controlling the fluctuation of training loss to less than 10%. Compared with the traditional fixed weight method, this adaptive mechanism can reduce redundant calculations while maintaining a high feature expression capability [6, 9].

In the lightweight design, the dynamic weight mechanism further adopts a channel-level gating strategy, as shown in Figure 3. The input features undergo lightweight convolution to generate channel-by-channel dynamic weights, which are then multiplied by the input features to obtain the final output. Due to the channel-by-channel operation, the number of parameters is only one percent of the backbone network, and the computational complexity is kept low. This mechanism has significant advantages during deployment, such as almost no additional latency, minimal memory usage, and the ability to be easily ported to a variety of hardware environments. This design ensures the efficiency of the model during training and inference [7, 9].

In the decoding phase, the model introduces a deformable convolution operation, which captures geometric deformations in the image by learning the offset of the convolution kernel. This method can effectively improve the segmentation performance of small objects and areas with blurred boundaries because the convolution sampling points can adaptively adjust their spatial position based on the input features. In order to avoid excessive increase in parameters, a grouped convolution structure is used to control the calculation, so that the increase in video memory remains within an acceptable range. Related research shows that deformable convolution has obvious advantages in improving the segmentation performance of complex structural regions [3, 4, 11].

During the inference phase, the hardware perception engine is further optimized through mixed-precision inference and layer fusion. The use of FP16 quantization not only significantly reduces video memory overhead but also maintains nearly the same accuracy as full precision. The layer fusion method reduces data transmission between computing units, thereby significantly shortening the inference time. Experimental results show that this inference engine stabilizes the single inference time to less than two seconds, which is nearly ten percent shorter than the baseline method, and can meet the real-time and high-efficiency requirements of medical imaging scenarios [8, 12, 15].

DOI: 10.54254/2755-2721/2026.TJ28958

4. Experimental results and discussion

Tablet 1. Quantitative results of different methods on liver segmentation

Test Method	Parameter quantity	Inference time	IoU value	GPU memory usage
3DU Net	20	53.29s	0.7509	3173.20MB
3DU Net+Deformable Convolution	20	52.42s	0.7524	3173.76MB
3DU Net+Dynamic Weight	20	53.11s	0.7515	3174.83MB
3DU Net+Deformable convolution + dynamic weights	20	48.80s	0.7744	3174.51MB

As shown in Table 1, this paper compares the baseline model 3DUNet and its different variants after adding deformable convolution and dynamic weights. The results show that after introducing deformable convolution alone, the inference time is reduced from 53.29 seconds to 52.42 seconds, and the IoU value is increased from 0.7509 to 0.7524. This shows that deformable convolution can effectively enhance the spatial modeling ability of the convolution kernel, so that the model can better adapt to the morphological changes of the target organ during feature extraction, thereby achieving better segmentation performance [2, 3, 12]. When dynamic weights are introduced alone, the IoU value is 0.7515, which is slightly improved compared with the baseline, while the inference time is only reduced from 53.29 seconds to 53.11 seconds, indicating that dynamic weights have a positive effect on accuracy, but their acceleration effect is limited when used alone.

In contrast, when deformable convolution and dynamic weights are combined, the model performs best. Specifically, the inference time is shortened to 48.80 seconds, which is about 8.4% less than the baseline, and the IoU value reaches 0.7744, which is about 3.1% higher than the baseline. This result shows that the two methods can play a synergistic role when combined. Deformable convolution improves the spatial adaptability of features, while the dynamic weight mechanism dynamically allocates weights according to input features, thereby enhancing the expression ability of information [4, 5, 9]. The combination of the two enables the model to achieve a double improvement in speed and accuracy while maintaining the same number of parameters and memory usage, verifying the effectiveness and practicality of the proposed method.

From the comparison of parameter size and memory usage, the four methods all maintain a memory overhead of approximately 20M parameters and 3173MB, with minimal differences. This result indicates that the proposed method does not introduce a significant computational burden in the model structure and is a lightweight improvement approach suitable for deployment in resource-constrained environments. This is consistent with the trend emphasized in existing research on controlling model complexity while ensuring accuracy [6, 8, 11].

Further analysis of the ablation experiment results shows that the improvement of a single module is limited, indicating that deformable convolution and dynamic weight mechanisms each have certain limitations. For example, while deformable convolution improves adaptability to complex boundaries, it is still insufficient in modeling long-range dependencies. Dynamic weighting mechanisms can achieve feature selection, but their effectiveness in modeling feature space is limited because they rely on the expressive power of convolution operations. The combination of the two complements each other's shortcomings to a certain extent, improving the overall robustness and generalization performance of the model.

DOI: 10.54254/2755-2721/2026.TJ28958

In addition, from the perspective of inference time, the proposed method achieves faster inference while improving accuracy, which is of great value for clinical applications. Medical image analysis often faces the challenge of processing large-scale three-dimensional data, while doctors have high requirements for real-time performance when planning surgery or screening diseases. The method proposed in this paper achieves faster reasoning and higher accuracy without adding additional hardware overhead, indicating that it has good application potential [1, 7].

Of course, this study still has certain limitations. For one thing, the experiments were primarily based on the LiTS dataset, which has a relatively uniform data distribution. Future validation is necessary on more cross-modal and cross-device datasets. Furthermore, this method still requires a long convergence time during training. Future research could explore combining it with more efficient optimization strategies to further reduce training overhead. Future research could also explore combining it with the Transformer architecture or multimodal fusion methods to further enhance the model's adaptability to complex medical scenarios.

5. Conclusion

This paper proposes an improved method based on a 3D U-Net to address the problems existing in medical image segmentation tasks, such as high computational overhead, low reasoning efficiency, and insufficient accuracy in small target segmentation. During the research process, a dynamic weight mechanism was introduced on the encoding side so that the model can adaptively adjust the weight distribution according to differences in input features, thereby reducing redundant calculations and improving the utilization of effective features. At the same time, the introduction of deformable convolutions at the decoding end enables the network to better capture subtle differences in spatial geometry, especially in the segmentation tasks of complex boundaries and small-scale targets, showing stronger adaptability. In the end, the two mechanisms were integrated to form a collaborative optimization solution that balances lightweight and high efficiency.

Experimental results show that this method achieves the dual improvements of shortening inference time and improving segmentation accuracy while keeping the video memory usage and parameter scale basically unchanged. Compared with the baseline model, introducing deformable convolution or dynamic weight alone can bring about certain improvements, but the combination of the two has the most superior performance, further verifying the effectiveness and innovation of collaborative design. This result shows that in lightweight medical image segmentation, an ideal balance can be achieved between speed, accuracy, and memory.

In summary, this study not only proposes a feasible dynamic-deformable collaborative optimization idea in theory but also demonstrates its application potential in the field of medical image segmentation in practice. Future work can be carried out in two directions. On the one hand, the method is extended to more complex clinical scenarios, such as multi-organ joint segmentation or lesion detection, and on the other hand, optimization methods combined with hardware-aware reasoning are further explored to promote its application in edge devices and real-time diagnosis.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Perslev, M., Dam, E. B., Pai, A., & Igel, C. (2019, October). One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. In International conference on medical image

DOI: 10.54254/2755-2721/2026.TJ28958

computing and computer-assisted intervention (pp. 30-38). Cham: Springer International Publishing.
[2] Li, Y., Hu, M., & Yang, X. MGDC-UNet: Multi-group Deformable Convolution for Medical Image Segmentation.
[3] Wang, L., Huang, J., Xing, X., & Yang, G. (2023). Swin deformable attention hybrid U-Net for medical image segmentation. arXiv preprint arXiv: 2302.14450.
[4] Azad, R., Niggemeier, L., Hüttemann, M., Kazerouni, A., Aghdam, E. K., Velichko, Y., ... & Merhof, D. (2024). Beyond self-attention: Deformable large kernel attention for medical image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1287-1297).
[5] Ou, Y., Chen, Y., Zhang, Y., Yang, S., Zhang, X., Zhou, Y., ... & Peng, L. (2025). Enhanced medical image segmentation via deep dynamic self-adjusting U-Net with multi-scale attention and semantic mitigation. The Visual Computer, 1-17.
[6] Zhao, Y., Li, J., & Liu, Y. (2024). Dynamic weight HiLo attention network for medical image multiple organ segmentation. International Journal of Imaging Systems and Technology, 34(1), e22966.
[7] Dar, M. F., & Ganivada, A. (2025, July). Dynamic Weight-Adjusted Ensemble Loss for Enhanced Medical Image. In Proceedings of Fourth International Conference on Computing and Communication Networks: ICCCN 2024, Volume 4 (Vol. 1292, p. 1). Springer Nature.
[8] Wang, Y., Wang, H., & Zhang, F. (2025). A Medical image segmentation model with auto-dynamic convolution and location attention mechanism. Computer Methods and Programs in Biomedicine, 261, 108593.
[9] Wang, X., Yu, J., Zhang, B., Huang, X., Shen, X., & Xia, M. (2025). LightAWNet: Lightweight adaptive weighting network based on dynamic convolutions for medical image segmentation. Journal of Applied Clinical Medical Physics, 26(2), e14584.
[10] Lu, Z., She, C., Wang, W., & Huang, Q. (2024). LM-Net: A light-weight and multi-scale network for medical image segmentation. Computers in Biology and Medicine, 168, 107717.
[11] Zhao, H., Yang, T., Xu, M., & Fu, Y. (2024, December). Deformable Transformer for 3D Medical Image Segmentation. In International Conference on Brain Inspired Cognitive Systems (pp. 126-135). Singapore: Springer Nature Singapore.
[12] Lee, H. H., Liu, Q., Yang, Q., Yu, X., Bao, S., Huo, Y., & Landman, B. A. (2023). Deformux-net: Exploring a 3d foundation backbone for medical image segmentation with depthwise deformable convolution. arXiv preprint arXiv: 2310.00199.
[13] Zhao, F., Wu, Z., Wang, L., Lin, W., Gilmore, J. H., Xia, S., ... & Li, G. (2021). Spherical deformable u-net: Application to cortical surface parcellation and development prediction. IEEE transactions on medical imaging, 40(4), 1217-1228.
[14] Zhao, C., Shuai, R., Ma, L., Liu, W., & Wu, M. (2021). Segmentation of dermoscopy images based on deformable 3D convolution and ResU-NeXt++. Medical & biological engineering & computing, 59(9), 1815-1832.
[15] Zhang, J., Chen, X., Yang, B., Guan, Q., Chen, Q., Chen, J., ... & Xia, Y. (2025). Advances in attention mechanisms for medical image segmentation. Computer Science Review, 56, 100721.