IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 July 2024, accepted 30 July 2024, date of publication 2 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437652

RESEARCH ARTICLE

MulA-nnUNet: A Multi-Attention Enhanced nnUNet Framework for 3D Abdominal Multi-Organs Segmentation

JIASHUO DING[1], WEI NI[1], JIAHUI WAN[2], XIAOJUN DENG[1], AND LANJUN WAN[1]

[1]School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China
[2]College of Mechanical and Electrical Engineering, Hunan Agricultural University, Changsha 410125, China

Corresponding author: Wei Ni (niwei@hut.edu.cn)

This work was supported in part by Hunan Provincial Natural Science Foundation of China under Grant 2023JJ30217, in part by the Scientific Research Foundation of Hunan Provincial Education Department under Grant 22A0408, in part by the Research Project on the Teaching Reform of Colleges and Universities in Hunan Province under Grant HNJG-20230752, in part by the Degree and Postgraduate Education Reform Project of Hunan University of Technology under Grant JGYB23030, and in part by the Innovation and Entrepreneurship Training Program for College Students of Hunan University of Technology under Grant 2023130.

ABSTRACT In the domain of medical image segmentation, the nnUNet framework is highly respected for its excellent performance and wide range of applications. However, the inherent bias of locality and weight sharing introduced by the continuous convolutional operations currently used limits the network's performance in modeling long-term dependencies. Furthermore, in the process of implementing residual links, certain limitations are encountered due to the substantial semantic discrepancy between the encoder's output feature maps and the decoder's. These limitations are seen in the direct application of skip connections for feature fusion and gradient propagation, which are known to impact the model's convergence speed and overall performance. In this paper, a novel framework is presented, namely Multi-Attention nnUNet (MulA-nnUNet), which utilizes nnUNet as the foundational network structure and integrates two key attention mechanisms: large kernel convolutional attention (LKA) and pixel attention (PA). LKA is embedded within the deep encoder, maintaining the effectiveness of shallow feature extraction and enhancing the deep neural networks' ability to understand long-range spatial dependencies. At the same time, the semantic distinction between the encoder and decoder's output map of features is decreased by the PA module, which helps to improve the effect of skip connection feature fusion. The complexity of the model is reduced by replacing the standard convolutions in the encoder and decoder layers with depthwise separable convolutions (DS), which have fewer parameters. The effectiveness of the proposed framework is confirmed by a set of ablation experiments and comparison experiments with current state-of-the-art models on the computed tomography (CT) subset of the multimodal abdominal multi-organ segmentation dataset (AMOS), which includes 500 CT scans, with 350 scans for training, 75 for validation, and 75 for testing. MulA-nnUNet shows improvements of 1.1% in mean dice similarity coefficient (mDSC) and 1.52% in mean intersection over union (mIoU), while the baseline model requires 5 times the floating point operations (FLOPs) and over 7 times the parameters (Params). Additionally, it demonstrates superior accuracy in segmenting organs such as the liver, stomach, aorta, and pancreas, thereby enhancing the accuracy of 3D abdominal multi-organ image segmentation.

INDEX TERMS Abdominal multi-organ image segmentation, attention mechanism, deep learning, nnUNet.

I. INTRODUCTION

Among the main tasks in the processing of medical images is regarded as medical image segmentation. In clinical practice,

The associate editor coordinating the review of this manuscript and approving it for publication was Sandra Costanzo.

manual or semi-manual segmentation technology is usually used for medical image segmentation, requiring appropriate clinical expertise to obtain clinically relevant contours, which is found to be very time-consuming. It is an indispensable tool in clinical diagnosis, treatment planning, and disease monitoring. High-resolution and rich 3D anatomical structure information for image segmentation is provided by CT technology, which is an important part of the field of medical imaging. However, as most deep neural network techniques are intended for 2D pictures, the correlation information in 3D structures is not adequately captured by the conventional 2D convolutional neural network (CNN). The model's performance deteriorates as a result of information loss. In addition, the detailed segmentation of multiple abdominal organs in CT is acknowledged as a difficult problem for automatic segmentation techniques as well as manual annotation, because the structural morphology of multiple abdominal organs is complex, and huge differences exist between different subjects and within the same subject, and the soft tissue contrast in the image is low [1].

U-Net [2] is acknowledged as a successful encoder-decoder baseline network, wherein the encoder component functions in a manner akin to conventional classification CNNs, aggregating semantic information through successive convolution operations at the price of reducing spatial information. By accepting semantic data from the bottom of the "U" and merging it with a higher-resolution feature map that is directly retrieved from the encoder via residual links, the decoder recovers the lost spatial information. Table 1 provides an overview of the strengths and weaknesses of various previous approaches extending the U-Net architecture. Among the many remarkable architectures, top performance in multiple medical imaging segmentation competitions has been achieved by the adaptive framework nnUNet [3]. A common network architecture similar to U-Net is used by nnUNet, but the complex manual method configuration process is systemized into fixed parameters, regular parameters based on dataset properties, and a minimum of empirical parameters for optimization. However, the richness of spatial information is gradually reduced during processing by continuously using convolution operations in nnUNet, resulting in the loss of critical connections between some distant pixels. Therefore, limitations are shown by it in capturing long-distance dependencies. Qurri and Almekkawy [4] noticed that CNN-based architecture has inherent biases that make it limited in simulating long-range dependencies. To collect long-distance context representations, a transformer is implemented at the bottleneck as a link between the encoder and the decoder. A three-level attention (TLA) module is built at the decoder layer, reducing the semantic gap between codecs by more precisely capturing local semantic representations. However, certain limitations are faced by the attention mechanism when compared to convolution operations. Adaptation to the spatial dimension is exclusively focused on, while adaptation in the channel dimension, which is crucial for visual tasks, is neglected. Murugesan et al. [5] trained a residual 3D U-Net by adding a residual path with a 3 × 3 × 3 layer of convolution and a layer of normalization at every encoding stage. The model is trained five times utilizing the dice similarity coefficient (DSC) loss function in conjunction with weighted cross-entropy loss. This method enables the concurrent segmentation of all 15 abdominal organs by employing magnetic resonance imaging (MRI)/CT scans, resulting in the achievement of generalized and excellent performance in segmentation for both CT and MRI cross-domain picture modalities. A linear attention mechanism known as LKA was proposed by Guo et al. [6]. It breaks down a 13 × 13 convolution into three different forms: a 5 × 5 profound convolution, a 5 × 5 in-depth dilated convolution with a dilation factor of 3, and a point-wise convolutive operation. This process incorporates the benefits of self-attention, including adaptability and long-term dependence. Furthermore, the advantages of convolution, such as the exploitation of local context information, are benefited from. In an effort to address the issue of precise brain tumor segmentation, Li et al. [7] experimented with the combination of 3D LKA and U-Net. Results demonstrate that LKAU-Net performs better in characterizing all three tumor subregions after training and assessing the multimodal brain tumor segmentation challenge (BraTS) 2020 dataset. There are documented average dice scores of 79.01%, 91.31%, and 85%, in that order. The whole tumor (WT), the enhancing tumor (ET), and the tumor core (TC) have Hausdorff distances of 26.27, 4.56, and 5.87 for 75%, 95%, and correspondingly. The VGA-Net proposed by Jalali et al. [8] integrates graph convolutional networks (GCNs) with attention mechanisms, effectively capturing the global structure of retinal vessels and maintaining segmentation continuity. This ensures an accurate representation of pixel-level information and structural details during the segmentation process. A concise and effective PA, similar in expression to spatial attention and channel attention but generating a 3D attention map as opposed to a 2D or 1D vector, is proposed by Zhao et al. [9], achieving good performance on lightweight networks. Shah et al. [10] offered EMED-UNet, which utilizes multiple U-Net structures with each U-Net gathering information in a separate receptive field. Several segmented outputs and receptive fields (one corresponding to each U-Net) are learned by the model to finally gather information. Deep supervision techniques are also introduced by the architecture to realize collaborative learning among multiple U-Nets. Fateh et al. [11] proposed the MRA-UNet for multilingual handwritten digit recognition, which used transfer learning to reduce the computational cost and maintain the quality of the enhanced image and the recognition accuracy of the model, solving the problem that handwritten digits in different languages have significant distance representation in the latent space. MA-UNet [12] employs compact attention U-Net as its foundational network structure, and by combining the characteristics produced by several intermediary layers for prediction, a multi-scale

TABLE 1. Overview of previous approaches.

Main methods	Strengths	Weaknesses	Authors
U-Net	Efficient training, end-to-end learning	Memory intensive and overfitting	Ronneberger et al. [2]
nnUNet	Self-Configuring and robust performance	Loss of long-distance spatial relationships	Isensee et al. [3]
TLA-integrated UNet++	Better localization capability and boundary delineation	Lack of channel adaptability	Qurri et al. [4]
Residual 3D U-Net	Broadly generalized performance	Longer training times and increased complexity	Murugesan et al. [5]
VAN	Better adaptability and long-term dependence	Large-scale self-supervised learning challenges	Guo et al. [6]
LKA-UNet	Improved tumors segmentation accuracy	Resource intensive	Li et al. [7]
VGA-Net	Pixel-level and Graph-level information, greater accuracy	Higher computational	Jalali et al. [8]
PAN	Lightweight design	Limited complexity handling	Zhao et al. [9]
EMED-UNet	Lightweight and accurate model, more feasible to deploy	Implementation complexity	Shah et al. [10]
MRA-UNet	High recognition accuracy, multilingual handwritten numeral recognition	Time-intensive	Fateh et al. [11]
MA-UNet	Multiscale method	Potential overhead	Cai et al. [12]
Efficient nnUNet	Reduced computational complexity	Potential accuracy trade-off	Magadza et al. [13]
Channel-Spatial-Attention-nnUNet	Enhanced feature relevance	Increased model complexity and computational overhead	McConnell et al. [14]

Abbreviations: TLA-integrated UNet++, UNet++ with Three-Level Attention; VAN, Visual Attention Network; LKAU-Net, Large Kernel Attention UNet; VGA-Net, Vessel Graph-based Attentional UNet for retinal vessel segmentation; PAN, Pixel Attention Network; EMED-UNet, Efficient Multi-Encoder-Decoder based UNet; MRA-UNet, Multi-Resolution Attention UNet; MA-UNet, Multiscale and Attention mechanism UNet for semantic segmentation of medical images.

method is used to achieve the integration and exploitation of global data at various stages. Furthermore, to better explore the global context information, the association mechanism between features and attention is built, and the attention mechanism is incorporated to concurrently articulate dependencies across both spatial and channel dimensions.

The following is a summary of this paper's primary contributions:

- In order to enhance the semantic segmentation capabilities of 3D abdominal multi-organ images, a new framework, MulA-nnUNet, is proposed in this paper, by utilizing nnUNet as the fundamental network structure while introducing the LKA and PA modules. Additionally, the standard convolution in the encoder and decoder layers is replaced by DS with fewer parameters, which obtains better segmentation results.
- To enhance the capacity of the deep neural network to grasp long-distance spatial relationships, the LKA module is introduced to the encoder's profound layer. The representation ability of the feature map is improved by this method, and the effective integration of features at different levels is promoted without affecting the efficiency of shallow feature extraction.
- The PA module is introduced in this paper to lessen the disparity in semantics across feature mappings transmitted through skip connections between the codecs. This method yields an improved representation of important areas within the feature maps while minimizing the impact of irrelevant or noisy regions, thereby improving the effectiveness of feature fusion.
- This model is put through a battery of ablation tests and contrasted with more sophisticated models that have been released recently. The efficiency of this strategy is confirmed by the experimental findings, which show that this model delivers superior outcomes than earlier models.

II. RELATED WORK

A. NNUNET AND ITS VARIANTS

In a conventional CNN, local characteristics are extracted from pictures using convolutional and pooling layers, and then the images are classified using fully connected layers. However, classic CNN frequently fails to satisfy the requirements of high-precision segmentation in the abdominal multi-organ segmentation assignment because of the image's complexity and the correlation of numerous organs. A deep learning framework created especially for segmenting images is called U-Net. Its unique codec architecture and residual links allow the network to effectively restore the spatial details of the image while preserving high-level semantic information and finally generate segmentation masks through an output layer for binary segmentation. In the disciplines of biomedical image segmentation and other areas, U-Net has demonstrated outstanding achievements and has become a classic model in image segmentation tasks. nnUNet is a U-Net-based deep learning framework that features adaptive setup and rule parameters to make it more user-friendly and appropriate for various medical image partitioning challenges [15]. By extracting the dataset fingerprint of dataset attributes, nnUNet models the interdependence of parameters. A set of heuristic rules is adopted to manipulate these dependencies. In this way, it can infer rule-based parameters to train up to three configurations using 5-fold cross-validation. nnUNet autonomously identifies the optimal combination of these models and decides whether post-processing is necessary. The robustness of nnUNet is not the result of a novel training plan, loss function, or network design; rather, it is the result of an intricate process of methodical manual method setup. nnUNet was originally developed on seven training datasets from the first phase of the medical segmentation decathlon challenge and won this competition [16]. In addition, in the automatic cardiac segmentation challenge (ACDC), nnUNet successfully segmented the dynamic magnetic resonance imaging (cine-MRI) heart images taken at two distinct time intervals. Accurate segmentation results of three parts of the heart are obtained. In this challenge, nnUNet ranked first in the open leaderboard. Isensee et al. [17] extended nnUNet by meticulously altering hyperparameters, including residual connections in the encoder, and creating a unique post-processing plan to compete in the AMOS 2022 competition. For task 1 (CT) and task 2 (CT+MRI), the final ensemble receives dice scores of 90.13% and 89.06%, respectively. Due to the significance of high-quality segmentation, most state-of-the-art models come at the expense of computational complexity. However, practical applications have a limited computational budget, so technical solutions that strike a balance between accuracy and available computational resources are needed. This is why Magadza and Viriri [13] extended the U-Net model in nnUNet. To reduce the number of network parameters and increase network efficiency, all standard convolutions are replaced with deep-separated convolutions. A bottleneck unit is also added to further reduce the number of parameters, and the skip connection uses a three-dimensional shuffle attention mechanism to enhance the network's segmentation performance. To prevent network deterioration, residual connections are also added. Using just 2.51 Mega (M) parameters and 55.26 Giga (G) FLOPs, the network obtained dice scores of 79.2%, 91.2%, and 84.8% for ET, WT, and TC improvement, respectively, on the BraTS 2020. Three types of attention mechanisms as well as additional ensemble mechanisms from advanced U-Net variations, such as residual, dense, and inception blocks, were added to the network architectural components of the nnUNet framework by McConnell et al. [14]. In addition, in Channel-Spatial-Attention-nnUNet, they integrated a variant of the newly proposed channel-attention mechanism, utilizing the channel attention block and spatial single attention block in turn to utilize channel and spatial attention, and added key modifications. This involves the substitution of the fully connected layer by a 1 × 1 × 1 convolutional layer, and the replacement of the addition operation following the fully connected layer with a concatenation operation prior to convolution. Such adjustments are made to enhance the preservation of information and to sustain the numerical differentiation between the maximum and average outputs for the layers of convolution that follow. The findings demonstrate that the application of attention variations to a tumor segmentation problem involving two or more target anatomical areas may effectively increase the segmentation performance and that the use of deep supervised structural features influences the segmentation performance.

B. LARGE KERNEL ATTENTION

The main goal of an attention mechanism in computer vision is to teach the system to disregard irrelevant input and to concentrate on important details. In recent years, this concept has been applied widely in various disciplines, including natural language processing [18], [19], speech recognition [20], [21], image processing [22], [23], and so forth. Building neural networks with attention mechanisms has become increasingly crucial as deep learning continues to progress swiftly. 3D abdominal images contain rich 3D structural information, and there are also complex relationships between various organs. Different organs exist in close proximity, resulting in unclear boundaries between organs. The perceptual range is increased by designing large kernel convolution operations to better grasp the image's long-range dependencies, which helps to process global structural information and channel adaptation. The implementation of 3D LKA was demonstrated in LKA-UNet [7], where the 3D LKA module was applied to the upsampling decoder layer, the attention map was created by utilizing sigmoid function activation, and 3D large kernel convolution was applied to the feature maps activated by group normalization and leaky ReLU. Before 3D large kernel convolution, the result is produced by multiplying this attention map by the feature map element by element. The final decoder layer compresses the feature channels into three using a $1 \times 1 \times 1$ convolution, and then creates three prediction probability maps using the sigmoid function. Sigmoid outputs are added to all except the two lowest layers for enhanced deep supervision and to facilitate gradient flow to preceding layers. The LKA module introduced by Guo et al. [6] in VAN is segmented into three components: spatial local convolution (deep convolution), spatial remote convolution (deep dilated convolution), and channel convolution (1×1 convolution). This segmentation aims to minimize computational demands and parameters while capturing distant relationships, thus enabling the estimation of a point's significance and the generation of the attention map.

C. PIXEL ATTENTION

The idea behind the pixel attention mechanism is to emphasize or suppress specific regions by calculating the weight of each pixel in the image and then applying these weights to the original pixel or feature map. With this method, the model may prioritize the areas that are more crucial for the current job and adaptively modify its attentional focus. This idea was first put forward and quickly refined in the field of natural language processing. It has since been applied to computer vision, particularly in tasks like object identification, segmentation, and image classification. Zhao et al. [9] applied PA to the lightweight convolutional neural network of image super-resolution (SR) and divided PA into two building blocks, namely, self-calibration block with pixel attention (SC-PA) and upsampling block with pixel attention (U-PA), which contained few parameters but could obtain good reconstruction effects. There is also a case of combining the PA module with U-Net [24], which is applied to the landslide recognition task. The PA module is utilized in each upsampling stage of the model, and a two-dimensional deconvolution layer is used to upsample the feature map, and then a connection function is used to connect the upsampling map with the corresponding encoder feature map and input into the PA module. The representation of key regions is further enhanced to help perform feature fusion.

III. METHOD

A. OVERVIEW OF MulA-nnUNet

The U-Net architecture serves as the foundation for nnUNet, where the convolutional layers have local receptive fields. Although the receptive field can be gradually expanded by multi-layer convolution and pooling, global contextual information in the image is still limitedly captured by this method. At present, the action of pooling in the downsampling procedure decreases the spatial resolution as the network layers deeper during the encoder stage, which impacts the network's ability to comprehend the global structure and long-distance dependencies within the image. There are still certain limitations on the overall comprehension of the global information, even when at the decoder stage the skip connection combines the context information from the high levels of the decoder with the precise information from the shallow layers of the encoder. The primary reason for this restriction is the feature map of the decoder stage is semantically different from the output of the encoder stage, making it difficult to effectively fuse different levels of features when direct skip connections are used, thereby impacting the segmentation performance.

In the proposed method, the input images undergo preprocessing inherited from the nnUNet framework before being fed into the MulA-nnUNet model. This model employs an attention mechanism to capture long-range dependencies in 3D medical images and bridges the semantic gap between the feature maps connected by skip connections in the encoder-decoder architecture.

Fig. 1 offers an overview of the workflow, while Fig. 2 elaborates on the network architecture generated by the MulA-nnUNet framework.

Firstly, the encoder's deep layer contains references to the LKA module, so that it is acted upon by the feature maps with large receptive fields. This enhancement of the network's overall performance is accomplished without compromising the effectiveness of shallow feature extraction, allowing for the more effective capture and utilization of long-range spatial information. For the output feature map of each encoder layer, pixel-level attention weighting is performed through the corresponding PA module before downsampling, highlighting important feature areas and enhancing the semantic information of key areas. This assists in the fusion of the skip connection with the upsampling feature map corresponding to the decoder stage, thereby reducing the semantic gap between them. Ultimately, the depthwise separable convolution takes the role of the regular convolution in the encoder and decoder layers, reducing both the number of parameters and the complexity of the model.

B. DESIGN OF THE LKA MODULE

As mentioned earlier, the network's capacity for feature extraction is improved by LKA, which provides a larger receptive field, effectively modeling long-range dependencies. The LKA module is implemented in the last three layers of the encoder. This is a result of the fact that each convolutional layer's receptive field grows as the network's depth rises. Therefore, at a deeper level, high-level and abstract feature maps with larger receptive fields can be acted upon by the LKA module. Moreover, by being applied to deeper layers, the number of feature maps that need to be processed is reduced through LKA, effectively capturing and exploiting long-range spatial information without significantly increasing the computational burden. The LKA module, seen in Fig. 3, expands the number of channels to $\frac{4}{3}C$ using a $1 \times 1 \times 1$ convolutional layer before performing gaussian error linear unit (GELU) activation, assuming that the number of channels is C. This allows the network to obtain a broad feature space, thereby capturing richer and more detailed information, which assists the model in understanding complex patterns in 3D abdominal multi-organ images. A large kernel convolution of $\frac{4}{3}C \times \frac{4}{3}C \times \frac{4}{3}C$

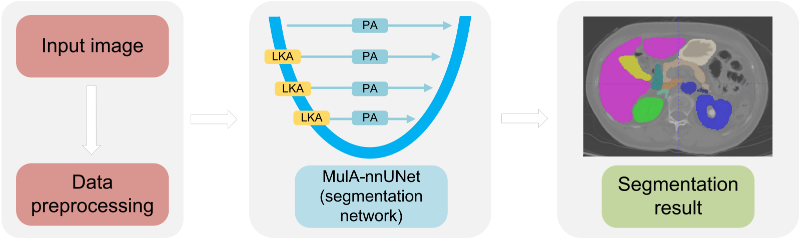


FIGURE 1. Workflow of the proposed method.

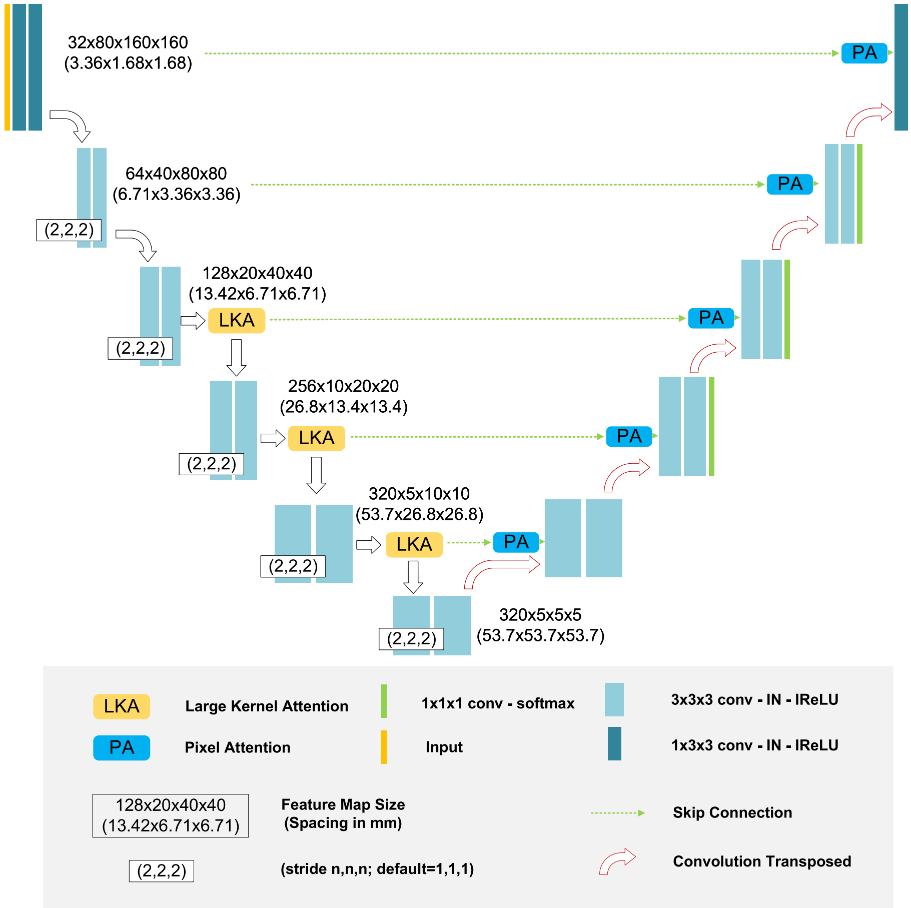


FIGURE 2. Network architecture generated by the MulA-nnUNet framework.

is decomposed into a depth-wise convolution (DW Conv) of size $(2d-1) \times (2d-1) \times (2d-1)$, a depth-wise dilated convolution (DWD Conv) of size $\left(\frac{4}{3}C \cdot \frac{1}{d}\right) \times \left(\frac{4}{3}C \cdot \frac{1}{d}\right) \times \left(\frac{4}{3}C \cdot \frac{1}{d}\right)$ with dilation (d, d, d), and a $1 \times 1 \times 1$ convolution $(1 \times 1 \times 1$ Conv). The attention map generated by the LKA module is then element-wise multiplied ($\otimes$) with the feature map previously activated by GELU. Ultimately, a $1 \times 1 \times 1$ convolutional layer is employed to restore the number of channels to their initial size, guaranteeing that the output of each residual connection retains the same channel dimension as the input. The entire LKA module can be written as follows:

$$E = \mathrm{GELU}\left(\mathrm{Conv}_{\mathrm{expansion}}\left(\mathrm{Input}\right)\right), \tag{1}$$

$$\mathrm{Atten} = \mathrm{Conv}_{1\times1\times1}\left(\mathrm{Conv}_{\mathrm{DW}}\left(\mathrm{Conv}_{\mathrm{DWD}}\left(E\right)\right)\right), \tag{2}$$

$$\mathrm{Output} = \mathrm{Input} + \mathrm{Conv}_{\mathrm{reduction}}\left(\mathrm{Atten} \otimes E\right). \tag{3}$$

where $E \in \mathbb{R}^{C\times D\times H\times W}$ displays the input feature map's feature map following the extension of the number of channels ($\mathrm{Conv}_{\mathrm{expansion}}$) and GELU activation function processing, $\mathrm{Atten} \in \mathbb{R}^{C\times D\times H\times W}$ represents the attention map. Atten and E are multiplied element-by-element. Following that, $\mathrm{Conv}_{\mathrm{reduction}}$ restores the number of channels, and the output

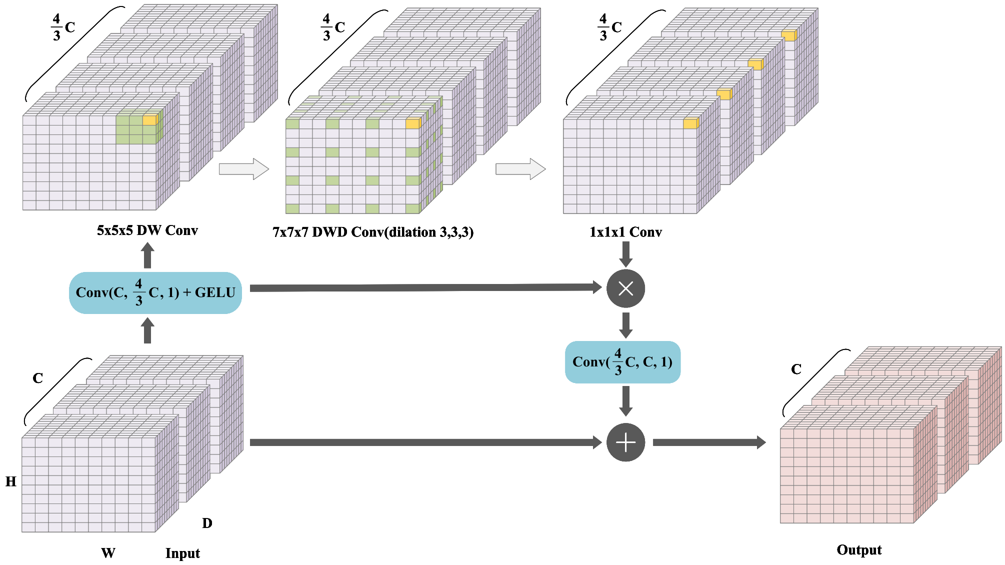

$\frac{4}{3}$C
$\frac{4}{3}$C
$\frac{4}{3}$C
5x5x5 DW Conv
7x7x7 DWD Conv(dilation 3,3,3)
1x1x1 Conv
Conv(C, $\frac{4}{3}$C, 1) + GELU
Conv($\frac{4}{3}$C, C, 1)
C
H
D
W
Input
C
Output

FIGURE 3. Large Kernel Attention (LKA) module: The input feature map is first processed through a 1 × 1 × 1 convolutional layer, which expands the number of channels to 4/3 times the original number of channels, and GELU activation is performed. Next, the feature map is processed by a 5 × 5 × 5 depth convolution (DW Conv), a 7 × 7 × 7 deep dilated convolution featuring a dilation factor of 3 (DWD Conv), and a 1 × 1 × 1 pointwise convolution to generate the attention map, which undergoes an element-wise multiplication with the feature map activated by gaussian error linear unit (GELU). The number of channels is restored to the original size by another 1 × 1 × 1 convolution. Ultimately, the final output feature map is constituted by adding the processed feature map to the original input feature map through a residual connection, culminating in the transformation process. In the figure, the location of the convolution kernel is represented by the colored grid, and the center point is represented by the yellow grid. The figure illustrates the DW Conv, DWD Conv, and 1 × 1 × 1 Conv, which are derived from the decomposition of a large kernel convolution with dimensions of 21 × 21 × 21. It presents merely a segment of the feature matrix resulting from this decomposition, specifically a corner, while excluding any representation of zero padding.

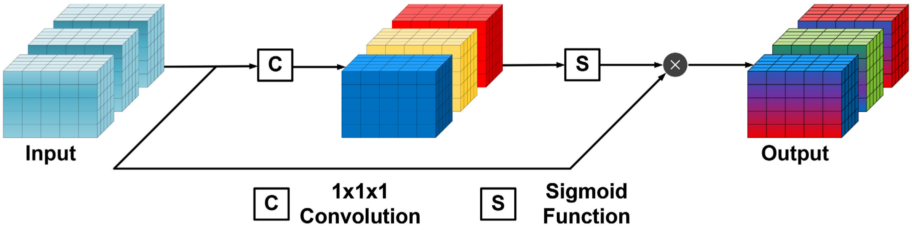


FIGURE 4. Pixel Attention (PA) module.

feature map is created by summing it with the input feature map.

C. DESIGN OF THE PA MODULE

In recent years, good performance has also been shown by the PA module by using the skip connection process. As previously pointed out, various restrictions in the feature fusion and gradient propagation of the direct skip connection are encountered, impacting the model's convergence speed and performance. This is because of the significant semantic gap between the encoder layer and the decoder layer's output feature maps. Therefore, the feature maps in skip connections can be weighed using PA to improve the feature representation of key semantics and reduce the interference of irrelevant features. The structure of the PA module is depicted in Fig. 4.

The input feature map Input $\in \mathbb{R}^{C\times D\times H\times W}$ in the encoder stage is convolved with 1 × 1 × 1, the features of each channel are transformed to discover the significance of every feature channel, and then the attention map Atten $\in \mathbb{R}^{C\times D\times H\times W}$ is generated by the sigmoid activation function. Ultimately, the input feature map and the attention map undergo element-by-element multiplication (⊗) to produce the output feature map, and the decoder layer's feature map is fused, which is described as follows:

$$\text{Atten} = \sigma_{\text{sigmoid}}\left(\text{Conv}_{1\times 1\times 1}\left(\text{Input}\right)\right), \tag{4}$$

$$\text{Output} = \text{Atten} \otimes \text{Input}. \tag{5}$$

D. LOSS FUNCTION

The hybrid loss function integrates cross-entropy (CE) loss and dice (DC) loss and is tuned for the model's segmentation and classification accuracy, leading to a superior segmentation effect and increased generalization capacity. The DC loss is derived from the dice coefficient, which is determined by the similarity between the model-predicted segmentation results and the true segmentation labels. It is calculated as follows:

$$\text{DCLoss} = -\frac{2\sum_{i=1}^{N} p_i g_i + \text{smooth}}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \text{smooth}}, \tag{6}$$

Here, p_i and g_i signify the predicted and actual values at the pixel i, correspondingly, with N symbolizing the aggregate pixel count, and smooth serving as a stabilization term introduced to preclude scenarios wherein the denominator approaches zero.

CE loss serves as a metric to assess the disparity between the model's predicted probability distribution and the actual distribution of the target, which is particularly effective for multi-class classification problems. It is calculated as follows:

$$\text{CELoss} = -\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right), \tag{7}$$

Here, $y_{o,c}$ is represented as a binary indicator denoting the membership of sample o in class c, whereas $p_{o,c}$ delineates the probability assigned by the model for sample o being part of class c. Here, M denotes the comprehensive count of classes, and log signifies the natural logarithm. The ultimate hybrid loss function is derived as delineated below:

$$\text{Loss} = \text{weight}_{\text{ce}} \cdot \text{CELoss} + \text{weight}_{\text{dice}} \cdot \text{DCLoss}. \tag{8}$$

Among them, $\text{weight}_{\text{ce}}$ and $\text{weight}_{\text{dice}}$ embody the respective contributions of CE loss and DC loss to the composite loss function. Through the strategic modulation of these weights and the synergistic integration of the merits of both CE loss and DC loss, an elevation in segmentation accuracy is achieved, concurrently with an augmentation in the model's sensitivity to the segmentation tasks. Since the loss is the sum of CE and DC losses, and the best CE loss is 0 while the best DC loss is -1, the overall best possible loss is -1.

IV. EXPERIMENT

A. DATASETS

A large and diverse abdominal multi-organ dataset containing 600 CT/MRI scans and over 74K annotated slices is provided by AMOS [25], featuring voxel-level annotations for 15 abdominal organs. The spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus are included, as shown in Fig. 5. Data are obtained from 600 patients diagnosed with abdominal tumors/abnormalities at Longgang District People's Hospital and Longgang District People's Hospital, Shenzhen, China. For the experiments, the AMOS-CT subset is used, where all 500 CT scans are interpolated to an isotropic voxel spacing of 1.0 mm × 1.0 mm × 1.0 mm.

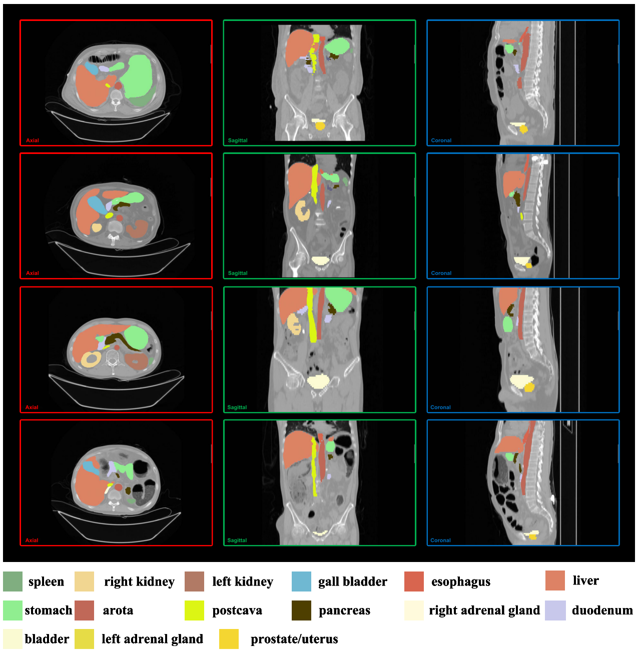


FIGURE 5. Examples of annotated slices in axial, sagittal, and coronal planes from the AMOS dataset.

B. EXPERIMENTAL DETAILS AND EVALUATION INDICATORS

1) EXPERIMENTAL DETAILS

In this experiment, Python 3 is utilized, and PaddlePaddle is chosen as the deep learning framework for training and testing on the NVIDIA A100 GPU. The details of hyperparameter configurations such as the optimizer, learning rate scheduler, and batch size will be determined after the hyperparameter selection ablation experiments. The experiments were conducted on the CT subset of the AMOS dataset, with a total of 400 epochs, which equates to 100,000 iterations. The DC_and_CE_loss function is adopted, and the evaluation metrics include intersection over union (IoU) and DSC. The experiments are conducted on the CT subset of the AMOS dataset, with a total of 400 epochs, equates to 100,000 iterations. Data augmentation operations from the nnUNet framework [16], such as gaussian noise, random rotation, random scaling, random elastic deformation, gamma-correction enhancement, and mirroring, are inherited by MulA-nnUNet. Additionally, MulA-nnUNet also inherits the data preprocessing operations [16] from the nnUNet framework, such as cropping, resampling, and normalization.

2) EVALUATION INDICATOR

In this experiment, two performance evaluation metrics are utilized, one being the DSC, which is commonly employed in medical image segmentation tasks, and the other being the

IoU. These metrics are used to assist in assessing the quality of different models. The DSC is extensively employed in the realm of medical image segmentation as a metric to ascertain the congruence between the prediction and the genuine label, thus facilitating the assessment of the model's efficacy. For a given organ i, where $\mathrm{Pre_i}$ denotes the pixel set corresponding to the predictive outcome and $\mathrm{Gt_i}$ represents the pixel set of the actual label, the formulation of the DSC is expressed as follows:

$$\mathrm{DSC_i} = \frac{2 \times |\mathrm{Pre_i} \cap \mathrm{Gt_i}|}{|\mathrm{Pre_i}| + |\mathrm{Gt_i}|}, \tag{9}$$

The IoU metric stands as the predominant measure for appraising the efficacy of models dedicated to semantic segmentation, referring to the coverage degree between the predicted region and the true annotation, namely, the ratio of their intersection region to the union region. For every specified organ i, wherein $\mathrm{Pre_i}$ is designated as the pixel ensemble of the forecasted outcome and $\mathrm{Gt_i}$ signifies the pixel ensemble of the authentic label, the IoU is expressed as follows:

$$\mathrm{IoU_i} = \frac{|\mathrm{Pre_i} \cap \mathrm{Gt_i}|}{|\mathrm{Pre_i} \cup \mathrm{Gt_i}|}. \tag{10}$$

C. HYPERPARAMETER OPTIMIZATION EXPERIMENTS

To identify the optimal hyperparameter configuration, a series of experiments is conducted focusing on three key factors using the MulA-nnUNet model: the optimizer (such as adaptive moment estimation with weight decay, AdamW [26]; adaptive moment estimation, Adam; and stochastic gradient descent, SGD), the learning rate scheduler (including polynomial decay, PD; cosine annealing decay, CAD; and step decay, SD), and the batch size. The evaluation metrics include the training loss and the mDSC, which is the average dice similarity coefficient for 15 abdominal multi-organs obtained from predictions on the validation set during training.

1) OPTIMIZER SELECTION EXPERIMENT

The performance of AdamW, Adam, and SGD optimizers is evaluated while keeping the learning rate scheduler and batch size constant. Fixed parameters are selected based on standard practices and preliminary experiments indicating stable training conditions.

- Optimizer evaluated: AdamW, Adam, SGD
- Fixed parameters:
  - Learning rate scheduler: Learning rate scheduler: PD is chosen for its ability to gradually decrease the learning rate throughout training, ensuring smooth convergence.
  - Batch size: 2. This batch size is chosen to provide stable gradient estimates, especially in situations where computational resources are limited.
- Result: Fig. 6 shows the training loss and mDSC trends for each optimizer evaluated. AdamW achieves a final training loss of -0.91 and an mDSC of 92.00%, demonstrating faster convergence and superior performance compared to Adam (training loss of -0.78 and mDSC of 78.33%) and SGD (training loss of -0.81 and mDSC of 74.77%). Due to its superior performance, AdamW is selected for subsequent experiments.

2) LEARNING RATE SCHEDULER SELECTION EXPERIMENT

This experiment assesses the impact of PD, CA, and SD learning rate schedulers on model performance.

- Learning rate schedulers evaluated: PD, CAD, SD
- Fixed parameters:
  - Optimizer: AdamW
  - Batch size: 2. To maintain consistency, a batch size of 2 is continued to be used.
- Result: Fig. 7 shows the training loss and mDSC trends for each learning rate scheduler evaluated. PD achieves a final training loss of -0.89 and an mDSC of 86.77%, demonstrating faster convergence and superior performance compared to CAD (training loss of -0.78 and mDSC of 76.63%) and SD (training loss of -0.71 and mDSC of 72.97%). Due to its superior performance, PD is selected for subsequent experiments.

3) BATCH SIZE SELECTION EXPERIMENT

This experiment explores the effect of different batch sizes (1, 2, 4) on model performance.

- Batch sizes evaluated: 1, 2, 4
- Fixed parameters:
  - Optimizer: AdamW
  - Learning rate scheduler: PD
- Result: Fig. 8 shows the training loss and mDSC trends for each batch size evaluated. Batch size 2 achieves a final training loss of -0.91 and an mDSC of 91.84%, demonstrating faster convergence and superior performance compared to batch size 1 (training loss of -0.83 and mDSC of 85.88%) and batch size 4 (training loss of -0.88 and mDSC of 85.67%). Due to its superior performance, batch size 2 is selected for subsequent experiments.

Based on the results of the hyperparameter optimization experiments, the best configuration for training the model is using the AdamW optimizer, the PD strategy for the learning rate scheduler, and a batch size of 2. This configuration ensures the optimal balance between training efficiency and model performance.

D. NETWORK COMPONENT ABLATION EXPERIMENTS

To verify the effectiveness of the individual components within the MulA-nnUNet framework, a series of ablation studies are conducted. The effects of the LKA module, PA module, and DS on the semantic segmentation performance of 3D abdominal multi-organ images are mainly focused on in these ablation experiments. By gradually

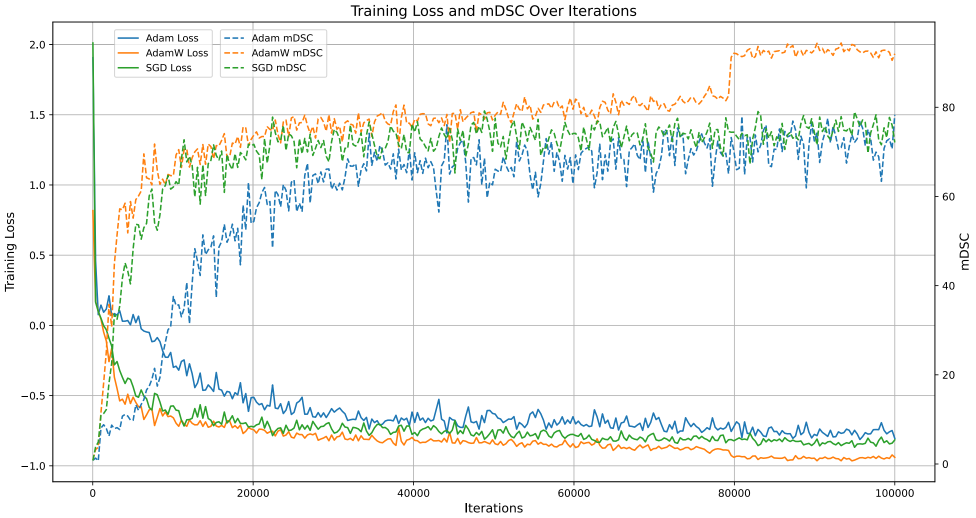


FIGURE 6. Result of optimizer selection experiment.

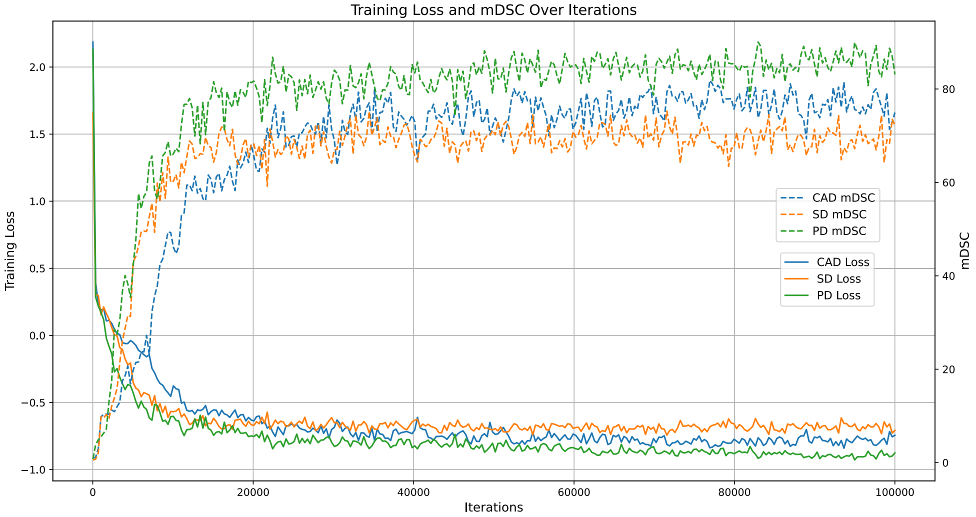


FIGURE 7. Result of learning rate scheduler selection experiment.

adding these modules on top of the baseline model, nnUNet, the contribution of each module to the overall performance can be identified. The specific model configuration is as follows:

- BL: Baseline model nnUNet, which does not contain any additional modules.
- BL+LKA: The LKA module is introduced in the last 3 layers of the encoder of the baseline model.
- BL+PA: The PA module is introduced in the skip connection of the baseline model.
- BL+LKA+PA+DS: Based on the LKA and PA modules, the DS is further introduced.

Table 2 displays each configuration's experimental results. The experimental findings demonstrate that the model gradually introduces the LKA and PA modules and replaces the usual convolution with DS to obtain certain performance gains on the 3D abdominal multi-organ image segmentation challenge. Specifically, an improvement of 0.84% in the mDSC and 1.19% in the mIoU is observed in the model by the incorporation of the LKA module. By adding the PA module, improvements of 0.34% in the mDSC and 0.64% in the mIoU are observed in the model, respectively. Among them, the highest values in mDSC and mIoU are achieved by the BL+LKA+PA configuration, showing an increase

TABLE 2. Results of ablation studies for the MulA-nnUNet framework, where the best values are shown in bold.

Models	mDSC (%)	mIoU (%)	FLOPs (G)	Params. (M)
Baseline (BL)	88.83	80.11	522.182	30.796
BL+LKA	89.67	81.30	529.212	32.241
BL+PA	89.17	80.75	526.252	30.987
BL+LKA+PA	**90.21**	**82.88**	531.753	32.228
BL+LKA+PA+DS	89.92	81.63	**101.487**	**4.088**

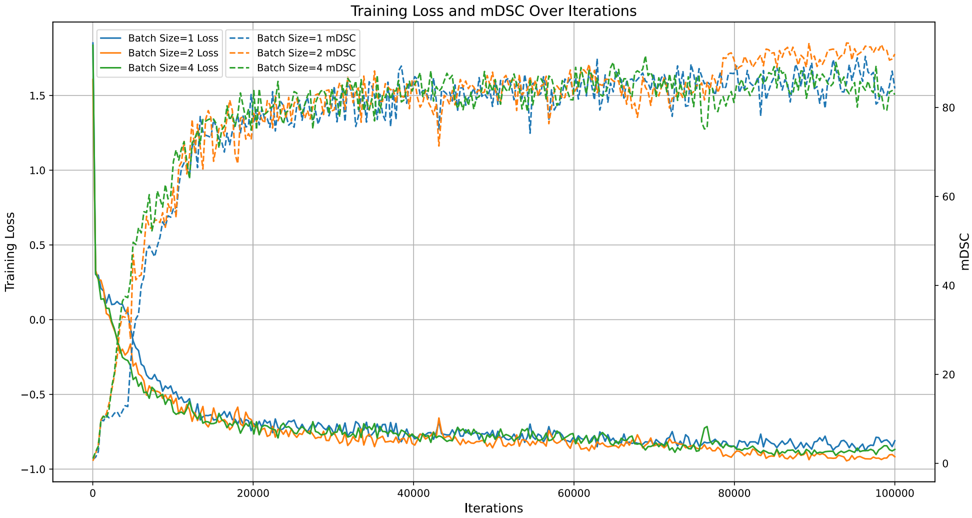


FIGURE 8. Result of batch size selection experiment.

TABLE 3. Evaluation of MulA-nnUNet against current leading-edge segmentation techniques on the CT subset of the AMOS dataset, as measured by the DSC. Note: spleen (Spl.), right kidney (Rki.), left kidney (Lki.), gallbladder (Gbl.), esophagus (Eso.), liver (Liv.), stomach (Sto.), aorta (Aor.), inferior vena cava (Ivc.), pancreas (Pan.), right adrenal gland (Rag), left adrenal gland (Lag.), duodenum (Duo.), bladder (Bla.), prostate/uterus (Pro./Ute.). Bold text indicates the best results.

Models	DSC (%)															
	Spl.	Rki.	Lki.	Gbl.	Eso.	Liv.	Sto.	Aor.	Ivc.	Pan.	Rag.	Lag.	Duo.	Bla.	Pro./Ute.	Avg.
nnUNet [3]	96.43	**96.53**	95.51	84.54	**87.21**	97.64	95.78	94.33	**87.49**	83.63	**75.03**	78.54	83.15	91.41	85.32	88.83
VNet [27]	94.28	92.12	92.46	73.80	79.31	94.58	82.64	91.96	83.28	80.32	71.56	73.29	71.84	79.68	67.59	81.91
nnFormer [28]	95.97	93.65	94.56	78.33	82.89	95.89	89.63	93.87	88.26	85.21	75.30	76.02	78.53	84.80	74.63	85.84
TransUNet [29]	96.12	93.01	94.78	77.88	82.05	94.97	89.64	93.60	86.57	82.45	75.82	76.37	79.56	83.98	73.78	85.37
SwinUNet [30]	95.86	92.37	94.05	76.98	81.69	94.12	88.87	92.91	86.94	82.00	75.81	76.36	79.16	83.20	72.98	84.89
UNETR [31]	92.78	87.48	90.53	66.84	73.61	94.12	78.98	92.99	82.46	74.12	68.13	65.31	62.54	77.91	67.56	78.36
MulA-nnUNet	**96.58**	95.73	**95.67**	**90.17**	85.49	**97.92**	**95.75**	**94.52**	86.98	**85.35**	74.36	**85.46**	**85.25**	**93.21**	**86.57**	**89.93**

TABLE 4. Evaluation of MulA-nnUNet against current leading-edge segmentation techniques on the CT subset of the AMOS dataset, as measured by the IoU. Note: Bold text indicates the best results.

Models	IoU (%)															
	Spl.	Rki.	Lki.	Gbl.	Eso.	Liv.	Sto.	Aor.	Ivc.	Pan.	Rag.	Lag.	Duo.	Bla.	Pro./Ute.	Avg.
nnUNet [3]	93.74	**91.62**	92.68	78.84	**71.32**	94.15	90.32	89.26	**78.20**	68.71	**53.02**	68.55	71.87	87.18	72.11	80.11
VNet [27]	90.11	88.62	89.65	69.77	66.42	93.21	80.10	87.65	73.59	67.35	50.73	65.92	66.04	73.91	48.99	74.14
nnFormer [28]	91.05	90.12	91.17	75.53	68.13	93.56	86.92	88.05	76.98	68.13	52.30	67.83	68.58	83.47	68.92	78.04
TransUNet [29]	90.73	89.52	90.57	75.14	67.70	92.63	86.79	87.70	76.48	67.44	51.52	67.42	68.23	83.22	68.10	77.55
SwinUNet [30]	90.12	89.15	89.72	74.93	67.36	92.49	86.69	86.94	75.69	66.89	50.91	67.37	68.12	82.24	67.87	77.10
UNETR [31]	86.34	82.23	85.77	68.87	62.63	87.18	79.43	82.75	69.97	61.11	45.32	63.14	63.70	73.56	45.68	70.51
MulA-nnUNet	**93.79**	91.34	**92.89**	**80.44**	71.63	**96.32**	**91.86**	**89.30**	77.86	**71.69**	52.97	**73.61**	**73.12**	**91.69**	**75.88**	**81.63**

of 1.38% and 2.77%, respectively. This demonstrates that these two attention mechanisms can effectively enhance the capture of long-distance spatial dependence and the prominence of important features, as well as promote the effective integration of features at various levels. Furthermore, the deployment of DS significantly diminishes both the computational complexity and the volume of parameters within the model. Although there is a marginal reduction in mDSC and mIoU, this outcome is deemed a tolerable compromise. The rationale behind this perspective lies in the balance between performance and efficiency, as vindicated by the notable enhancement in computational efficiency and the reduction of the model's complexity.

E. COMPARATIVE STUDY

To ascertain the impact of the introduced MulA-nnUNet architecture on the task of 3D abdominal multi-organ image segmentation, this section undertakes a comparative analysis between the proposed model and existing sophisticated models. Included in these models are nnUNet (Baseline) [3], VNet [27], nnFormer [28], TransUNet [29], SwinUNet [30], and UNETR [31]. Tables 3 and 4 display the related results,

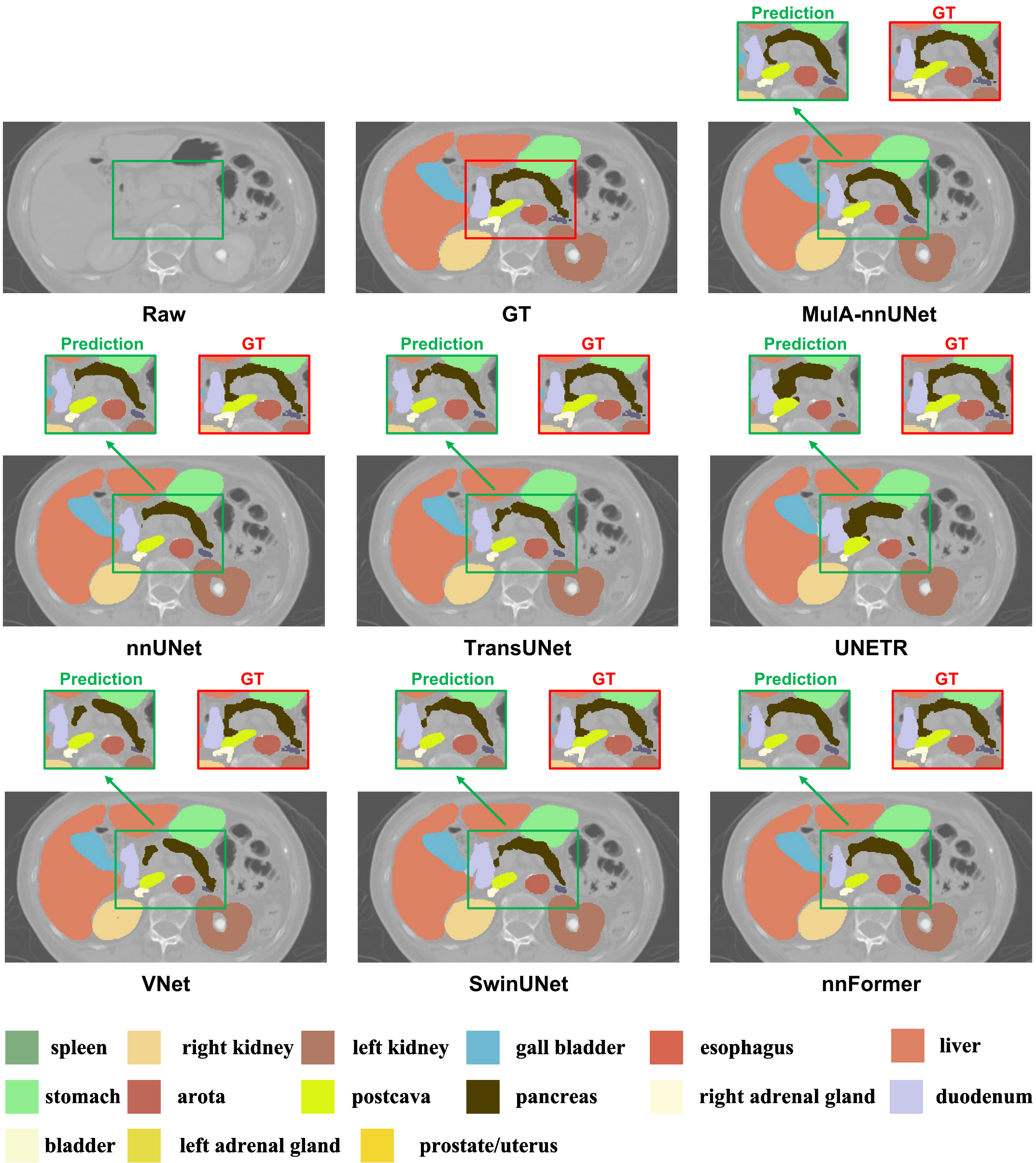


FIGURE 9. Visual comparison with other models on AMOS-CT.

TABLE 5. Comparison of performance metrics between MulA-nnUNet and various advanced models. Note: Bold text indicates the best results.

Models	FLOPs (G)	Params. (M)	Runtime (s)
nnUNet [3]	522.182	30.796	0.347
VNet [27]	887.477	45.674	**0.309**
nnFormer [28]	265.986	150.329	0.455
TransUNet [29]	128.473	105.283	0.549
SwinUNet [30]	106.542	27.171	0.459
UNETR [31]	286.347	93.436	0.337
MulA-nnUNet	**101.487**	**4.088**	0.315

whereas Fig. 9 provides instances of the corresponding segmentation.

Tables 3 and 4 describe the results of the experiments of MulA-nnUNet and various other models on the CT subset of the AMOS dataset. The mDSCs of nnUNet, VNet, nnFormer, TransUNet, SwinUNet, and UNETR are reported as 88.83%, 81.91%, 85.84%, 85.37%, 84.89%, and 78.36%, respectively. The mIoUs obtained are reported as 80.11%, 74.14%, 78.04%, 77.55%, 77.10%, and 70.51%, respectively. The highest mIoU and mDSC are achieved by MulA-nnUNet, reported as 89.93% and 81.63%, respectively. Increases in the mDSC of 1.1%, 8.02%, 4.09%, 4.56%, 5.04%, and 11.57%, respectively, are achieved by MulA-nnUNet contrasted with the results of the remaining six models. Increases in the mIoU of 1.52%, 7.49%, 3.59%, 4.08%, 4.53%, and 11.12%, respectively, are observed. From the perspective of each organ, the best performance on 11 abdominal organs is achieved by the proposed MulA-nnUNet, which especially shows great advantages in the segmentation of organs such as the liver, stomach, aorta, and pancreas. Better results than those of other models are achieved by our proposed model, with mDSC and mIoU scores that are 1.1% and 1.52% higher, respectively, than the second-best scoring model, nnUNet.

Table 5 describes the performance of MulA-nnUNet and various other advanced models in terms of runtime, model parameters, and FLOPs. Compared to the nnUNet model, it requires only 20% of the FLOPs and 13% of the parameters while maintaining a relatively fast runtime. Fig. 9 displays the visual comparison with various models on the AMOS-CT dataset. More accurate details in the segmentation map, better handling of close or intersecting organ boundaries, and accurate depiction of the contour and internal details of organs are achieved by the segmentation map predicted by MulA-nnUNet. It is confirmed that our suggested approach works well for enhancing the segmentation accuracy of 3D abdominal multi-organ images.

V. CONCLUSION

Medical image analysis is thought to include medical image segmentation as a crucial component. The U-Net architecture is carefully analyzed in this paper in an attempt to identify any possible areas for improvement. It is found that the continuous convolution operation leads to a gradual reduction of spatial information in the feature map with the downsampling, thus resulting in the loss of some important correlations between distant pixels. In this paper, LKA is implemented prior to down-sampling within the deep encoder portion of the network. This modification augments the deep neural network's capacity to apprehend spatial dependencies over extended distances, thereby enhancing the representational efficacy of the feature maps. Moreover, distinctions are observed between the features transmitted from the encoder network and those conveyed through the decoder network. To coordinate these two sets of incompatible features, the addition of PA between them is proposed to strengthen the representation of important regions in the feature map and reduce the influence of irrelevant or noisy regions, which helps enhance the two feature maps' integration impact. With the addition of LKA and PA modules, a significant increase in the model's computational complexity is observed, prompting the replacement of the standard convolution in the encoder/decoder layer by DS. This adjustment reduces the model's complexity while maintaining competitive segmentation performance. Ultimately, the experimental results on the CT subset of the AMOS dataset validate the efficacy of the suggested method.

In order to evaluate our method's universality and application, future work will entail testing and validating it on other datasets and medical imaging modalities, including positron emission tomography (PET) and MRI. Additionally, further exploration and enhancement of the feature fusion strategy will be conducted to improve the model's performance and robustness.

REFERENCES

[1] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Med. Image Anal.*, vol. 55, pp. 88–102, Jul. 2019.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[3] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[4] A. Al Qurri and M. Almekkawy, "Improved UNet with attention for medical image segmentation," *Sensors*, vol. 23, no. 20, p. 8589, Oct. 2023.

[5] G. K. Murugesan, D. McCrumb, E. Brunner, J. Kumar, R. Soni, V. Grigorash, A. Chang, A. Peck, J. VanOss, and S. Moore, "Automatic abdominal multi organ segmentation using residual UNet," *bioRxiv*, 2023, doi: 10.1101/2023.02.15.528755. [Online]. Available: https://www.biorxiv.org/content/early/2023/02/16/2023.02.15.528755

[6] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Dec. 2023.

[7] H. Li, Y. Nan, and G. Yang, "LKAU-Net: 3D large-kernel attention-based U-Net for automatic MRI brain tumor segmentation," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, vol. 13413. Cham, Switzerland: Springer, 2022, pp. 313–327.

[8] Y. Jalali, M. Fateh, and M. Rezvani, "VGA-Net: Vessel graph based attentional U-Net for retinal vessel segmentation," *IET Image Process.*, vol. 18, no. 8, pp. 2191–2213, Jun. 2024.

[9] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 12537, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 56–72.

[10] K. D. Shah, D. K. Patel, M. P. Thaker, H. A. Patel, M. J. Saikia, and B. J. Ranger, "EMED-UNet: An efficient multi-encoder–decoder based UNet for medical image segmentation," *IEEE Access*, vol. 11, pp. 95253–95266, 2023.

[11] A. Fateh, R. Tahmasbi Birgani, M. Fateh, and V. Abolghasemi, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," *IEEE Access*, vol. 12, pp. 41381–41395, 2024.

[12] Y. Cai and Y. Wang, "MA-UNet: An improved version of UNet based on multi-scale and attention mechanism for medical image segmentation," *Proc. SPIE*, vol. 12167, pp. 205–211, Mar. 2022.

[13] T. Magadza and S. Viriri, "Efficient nnU-net for brain tumor segmentation," *IEEE Access*, vol. 11, pp. 126386–126397, 2023.

[14] N. McConnell, N. Ndipenoch, Y. Cao, A. Miron, and Y. Li, "Exploring advanced architectural variations of nnUNet," *Neurocomputing*, vol. 560, Dec. 2023, Art. no. 126837.

[15] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," 2019, *arXiv:1904.08128*.

[16] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*.

[17] F. Isensee, C. Ulrich, T. Wald, and K. H. Maier-Hein, "Extending nnU-Net is all you need," in *Proc. BVM Workshop*. Wiesbaden, Germany: Springer, 2023, pp. 12–17.

[18] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, Oct. 2021.

[19] D. Hu, "An introductory survey on attention mechanisms in NLP problems," in *Proc. Intell. Syst. Conf. (IntelliSys)*, vol. 1038. Cham, Switzerland: Springer, 2019, pp. 432–448.

[20] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.

[21] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–14.

[22] S. Ghaffarian, J. Valente, M. van der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sens.*, vol. 13, no. 15, p. 2965, Jul. 2021.

[23] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[24] B. Pan and X. Shi, "Fusing ascending and descending time-series SAR images with dual-polarized pixel attention UNet for landslide recognition," *Remote Sens.*, vol. 15, no. 23, p. 5619, Dec. 2023.

[25] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan, and P. Luo, "AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36722–36732.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

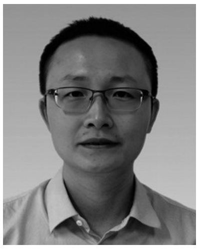
[28] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnFormer: Volumetric medical image segmentation via a 3D transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 4036–4045, 2023.

[29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[30] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13803. Cham, Switzerland: Springer, 2022, pp. 205–218.

[31] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1748–1758.

JIASHUO DING was born in Henan, China, in 2004. He is currently pursuing the B.S. degree in artificial intelligence with Hunan University of Technology, Zhuzhou, China. His research interests include medical image processing and optical character recognition.



WEI NI was born in Hunan, China, in 1981. He received the B.S. and Ph.D. degrees in communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2003 and 2012, respectively. He is currently an Assistant Professor with the School of Computer Science, Hunan University of Technology, Zhuzhou, China. He has published many research papers in international conferences and journals, such as IPEC, CCBR, and CSMA. His research interests include medical image processing, virtual and reality technology, and intelligent fault diagnosis.

JIAHUI WAN was born in Hunan, China, in 2005. She is currently pursuing the B.S. degree in robotics engineering with Hunan Agricultural University, Changsha, China. Her research interests include medical image processing and intelligent robots.

XIAOJUN DENG was born in Hunan, China, in 1974. He received the M.S. degree in computer science and technology from the National University of Defense Technology, Changsha, China, in 2004. He is currently a Full Professor with the School of Computer Science, Hunan University of Technology, Zhuzhou, China. He has published many research papers in international conferences and journals, such as *International Journal of Security and Networks*, *Journal of Computer Science and Engineering*, and IEEE ACCESS. His major research interests include industrial big data analysis, industry equipment health management, the Internet of Things, and image processing.

LANJUN WAN was born in Hunan, China, in 1982. He received the B.S. and M.S. degrees in computer science and technology from Hunan University of Technology, Zhuzhou, China, in 2005 and 2009, respectively, and the Ph.D. degree in circuits and systems from Hunan University, Changsha, China, in 2016. He is currently an Associate Professor with the School of Computer Science, Hunan University of Technology. He has published many research papers in international conferences and journals, such as *Knowledge-Based Systems*, IEEE SENSORS JOURNAL, *Measurement*, and *Journal of Parallel and Distributed Computing*. His research interests include industrial big data analysis, intelligent fault diagnosis, intelligent production scheduling, and high-performance computing.