Original Article

Dual-channel grouped cross-dimension attention V-Net for pulmonary nodule segmentation

Lihong Zhang^, Tong Liu, Yingbo Liang, Yuzhuo Li, Wenwu Zhang, Junding Sun

College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

Contributions: (I) Conception and design: L Zhang; (II) Administrative support: J Sun; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: L Zhang, T Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Junding Sun, PhD. College of Computer Science and Technology, Henan Polytechnic University, 2001 Century Avenue, Jiaozuo 454003, China. Email: sunjd@hpu.edu.cn.

Background: Accurate segmentation of pulmonary nodules is crucial for the diagnosis and treatment of early-stage lung cancer as it can aid clinicians in formulating effective treatment plans, increasing the chance of early detection and treatment, and reducing mortality. However, pulmonary nodules are similar to surrounding tissues, and the location, size, and quantity of nodules in different patients are unpredictable, posing challenges to accurate segmentation. This study aimed to develop a new deep learning network based on V-Net to address the deficiencies in pulmonary nodule segmentation tasks.

Methods: This work proposes a dual-channel grouped cross-dimension attention V-Net (DGCA V-Net) model for computed tomography (CT) pulmonary nodule segmentation. In downsampling, the model uses the global grouped coordinate attention (GGCA) module to comprehensively capture multidimensional global information and reduce the loss of feature information. In the bottom decoding path, the grouped split attention (GSA) module is adopted to effectively reduce the loss of detailed information. A dual-input guided feature aggregation (DGA) module is introduced between the encoding and decoding paths to effectively alleviate the impact of inaccurate localization learning on the detailed segmentation of pulmonary nodules.

Results: The proposed model was trained and evaluated with the Lung Nodule Analysis 2016 (LUNA16) pulmonary nodule public dataset. The segmentation performance was statistically analyzed with four indicators: the Dice similarity coefficient (DSC), intersection over union (IoU), precision, and recall. The proposed DGCA V-Net model significantly outperformed the baseline model V-Net in all metrics: the DSC increased by 4.72% to 0.7921, the IoU increased by 6.92% to 0.6662, the precision increased by 2.90% to 0.8102, and the recall increased by 4.80% to 0.7993.

Conclusions: Based on experimental data, the strategy presented in this study significantly improves lung nodule segmentation accuracy. The ablation tests also confirm the proposed module's strong segmentation and generalization abilities. This segmentation model is expected to be applied to other medical segmentations. Our solution is open-source and online (https://github.com/freshmancodes/Pulmonary-Nodule-Segmentation).

Keywords: Pulmonary nodule segmentation; V-Net; grouped coordinate attention (grouped CA); dual-input-guided feature aggregation (DGA)

Submitted Nov 04, 2024. Accepted for publication Jul 10, 2025. Published online Sep 18, 2025.
doi: 10.21037/qims-24-2434
View this article at: https://dx.doi.org/10.21037/qims-24-2434

^ ORCID: 0000-0003-4981-2555.

Introduction

Background

As a major threat to human health, lung cancer is among the tumor types with the fastest-rising mortality rate. In 2020, there were 9.96 million cancer-related deaths worldwide, with lung cancer accounting for 1.8 million (18.07%) of these deaths, placing it first among malignancies. Moreover, the 5-year survival of patients with lung cancer is merely 18.6%. Since many cases are detected in the middle or late stages due to the absence of clear early symptoms, the survival rate is less than 5%. Consequently, early diagnosis is crucial, and with prompt therapy, the 5-year survival rate can rise to 56% (1). Pulmonary nodules that are smaller than 30 mm in diameter are typically the first signs to appear (2). The primary screening method for lung cancer is computed tomography (CT); however, due to the volume of images it produces, the diagnosis process for radiologists is arduous and involves a high chance of error. In an effort to increase the early detection rate of lung cancer, it is imperative to create an accurate and efficient pulmonary nodule segmentation approach. The biggest contributor to segmentation difficulty is the fuzzy edge contour of pulmonary nodules, which resembles vascular tissue. This and other issues in the effective implementation of pulmonary nodule segmentation urgently need to be resolved.

CT scans of the lungs are three-dimensional (3D) data from images, yet due to limitations in processing power in lung nodule segmentation, they are often converted to two-dimensional (2D) images, which may lead to the loss of crucial spatial information. Therefore, this study used 3D image data as input and developed the dual-channel grouped cross-dimension attention V-Net (DGCA V-Net) based on the V-Net network for lung nodule segmentation so as to preserve the 3D spatial features of nodules and reduce information loss. The primary contributions of this study are as follows:

(I) The global grouped coordinate attention (GGCA) module is introduced during the downsampling phase. Channel weight information and multidimensional location information can be extracted by this module. This improves the model's capacity for generalization by taking into account channel information that is useful for pulmonary nodule segmentation from several dimensions. Moreover, it strengthens the correlation between features, aids the model in comprehending the links between channels, and reduces feature information loss by thoroughly capturing multidimensional global information.

(II) The grouped split attention (GSA) module is incorporated into the upsampling process. This module efficiently aggregates and reorganizes features, minimizing detail loss and increasing segmentation accuracy to return the deep feature map to a state that is similar to the original feature map. Furthermore, the GSA module improves the model's comprehension of intricate structures, attention to the interactions between feature groups, and the collection of contextual information.

(III) The dual-input-guided feature aggregation (DGA) module is added to the skip connection. This innovation effectively reduces the redundant information generated when high- and low-resolution features are combined, thereby improving the accuracy of nodule localization and improving the segmentation performance. The DGA module optimizes the information flow by intelligently integrating features from different resolutions to ensure that important features are fully displayed.

Related work

Two major categories can be used to classify lung nodule segmentation techniques. The common manual feature extraction techniques, including threshold segmentation (3) and region growing (4), are frequently influenced by texture fluctuations and background noise, leading to a less-than-ideal performance. Meanwhile, by training deep neural networks (5-7), deep learning-based automatic feature extraction techniques may successfully identify lung nodule features, greatly increasing segmentation accuracy. Ronneberger *et al.*'s U-Net architecture, frequently applied in the segmentation of medical images, efficiently captures contextual information by connecting the encoder and decoder with skip links, comprising an encoding path and a decoding path, respectively (8). However, its core architecture focuses largely on local feature extraction, which may lead to a restricted ability to collect global contextual information. YOLOv5 (9), Wavelet U-Net++ (10), 3D-DenseUNet (11), and other new models and variations of U-Net have since been proposed. Milletari *et al.* (12) introduced the V-Net model, which extracts features from

3D medical images using residual modules, an encoding path, and a decoding path. Despite its strong segmentation performance, the attention and segmentation effect of V-Net on the lesion area are still insufficient. Several novel segmentation techniques have been developed in an attempt to overcome this challenge. Liu *et al.* introduced a novel pulmonary nodule segmentation technique (13) that makes use of the residual edge improvement module to successfully improve edge features in the data while reducing redundant information. The 3D spatial convolutional pooling pyramid module enables the integration of multiscale features, and the 3D coordinate attention (CA) network improves the efficient propagation of spatial information in the encoding layer. However, this is limited by the problem of class imbalance in the dataset. A spatial and channel squeeze-and-excitation (scSE) network based on nonlocal blocks was proposed by Zhou *et al.* (14). In an effort to compensate for the inadequate spatial dependency in convolutional neural networks (CNNs), our proposed architecture adds nonlocal blocks at the bottom. The use of the scSE module greatly increases picture recognition accuracy. However, this strategy fails to fully learn the intricate interactions between features at various levels, leading to a low efficiency of information transfer. The enhanced residual U-Net was first proposed by Ji *et al.* (15). In this design, deep convolution and dense spatial pyramid pooling are used to improve feature extraction and multiscale information processing, and a channel and spatial attention mechanism is included in the decoder to maximize global pixel attention. However, these improvements reduce the potential for generalization by increasing the number of parameters and computing complexity and are prone to overfitting on short datasets. Xu *et al.* proposed a V-Net network (16) with a dual-branch feature module and the reverse attention context module, which improves the segmentation of target features at the jump connection, thus enhancing the texture features at the pulmonary nodule edge. However, if these two modules are introduced too frequently, the initially helpful edge information may be weakened due to the recurrent use of deep features. When faced with complicated pulmonary nodule structures, the model's segmentation accuracy may be diminished due to the repetitive fusing of features, which may result in an excessively smoothed effect in the recording of tiny edge changes. An improved V-Net segmentation network based on selective kernels was developed by Wang *et al.* (17). The general architecture of this network is based on the conventional V-Net concept. In this iteration, multiscale feature information can be efficiently recovered via the addition of selective convolution kernel with a soft attention mechanism to the selection kernel network. Due to this design, the model can handle inputs with varying resolutions more flexibly. Although this approach increases model complexity, the segmentation capacity is nonetheless enhanced. Dutande *et al.* (18) designed the U-Net model, which increases accuracy, particularly in the small-nodule segmentation test, by enhancing significant features and suppressing unimportant ones. To accomplish 2D segmentation, the model also incorporates a U-Net design based on channel attention. It should be noted that model disregards the depth information present in 3D data. An effective multiscale, fully convolutional U-Net model was proposed by Agnes *et al.* (19). This architecture greatly enhances the capacity to extract detailed information by utilizing multiscale convolution technology. However, the framework still lacks the ability to segment the features of pulmonary nodule boundaries and is unable to accurately depict the subtleties of nodule contour. Additionally, this model's segmentation effect is inadequate for adherent nodule types, which affects the segmentation performance as a whole. An enhanced U-Net-based feature extraction model was proposed by Li *et al.* (20). A spatial attention module and a feature improvement module are introduced in this framework to enable the network to extract more varied and effective information. Additionally, to acquire more detailed context information, the model integrates a multiscale module in the skip connection. Yet, the network's decoding path is inadequate for feature recovery, which could lead to the loss of crucial information during feature reconstruction. This flaw could impact the model's ability to handle intricate nodules.

Methods

With the widespread use of attention mechanisms in computer vision tasks in recent years, networks are now able to concentrate on more discriminative regions, improving feature representation and overall model performance. To better illustrate the rationale behind our model design, we briefly review several representative attention mechanisms. The selective kernel network (SK-Net) (21), involving a dynamic kernel selection mechanism, was developed to adaptively model multiscale features, enhancing the network's capacity to perceive various scales; meanwhile, squeeze-and-excitation network (SE-Net) (22) uses global pooling to model interchannel dependencies, thereby improving the expressiveness of channel attention.

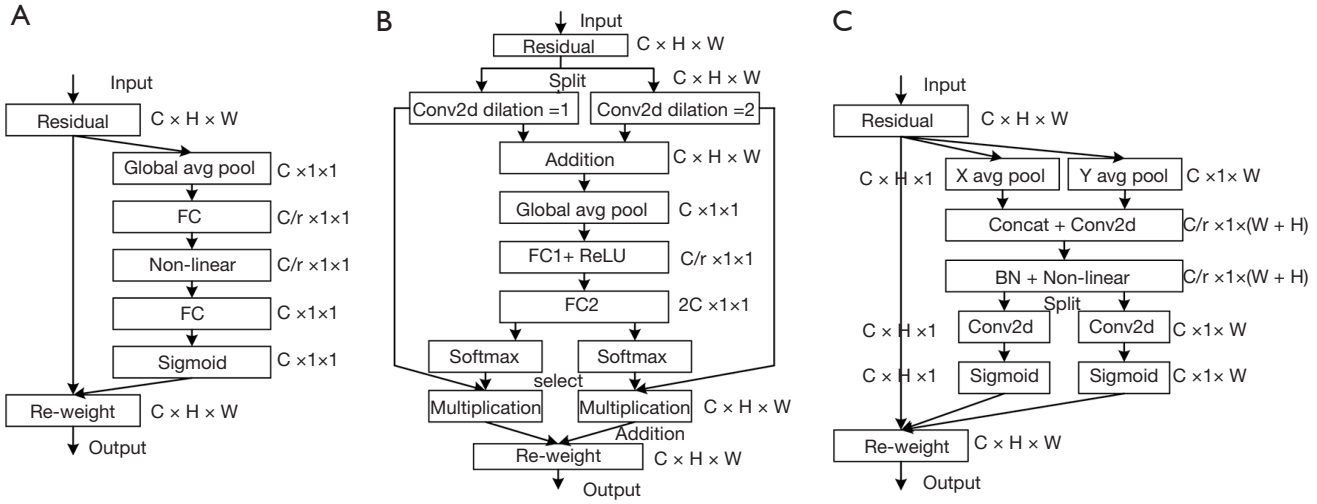

Figure 1 Structures of three representative attention modules. (A) The SE module, which models interchannel dependencies through global average pooling. (B) The SK module, which introduces dynamic kernels to adaptively select multiscale features. (C) The CA module, which integrates spatial positional information with channel attention to enhance feature representation. Avg pool, average pooling; C, channels; CA, coordinate attention; Conv, convolution; FC, fully connected; H, height; r, reduction ratio; ReLU, rectified linear unit; SE, squeeze and excitation; SK, selective kernel; W, width.

Moreover, CA (23) integrates positional information into the channel attention mechanism, balancing spatial structure and channel relationships, thereby enhancing performance in structural modeling tasks. Furthermore, the attention gate (AG) (24) mechanism has been widely used in medical image segmentation networks' skip connections. It dynamically filters feature paths to eliminate unnecessary information and improve the localization of important regions. *Figure 1* presents the structures and differences of the three current mainstream attention mechanisms.

The study's design was significantly influenced by the previously indicated attention mechanisms. They are less appropriate for modeling complex structures and achieving fine-grained segmentation in 3D medical images because the majority of them are based on 2D images and include drawbacks such as coarse granularity and inadequate redundancy control in feature fusion and spatial modeling. We suggest three attention modules based on a V-Net framework to address these problems, each of which is designed to address a distinct structural challenge: global 3D spatial modeling is improved by GGCA, feature reorganization and detail perception are improved by GSA during the decoding stage, and the fusion strategy of high- and low-resolution features in skip connections is optimized by DGA. These modules work together to provide a 3D segmentation network architecture that is more effective, precise, and structurally aware.

The general architecture of the DGCA V-Net module for lung nodule segmentation and the structures and functions for each of the GGCA, GSA, and DGA modules are described in detail in the following sections.

Overall architecture

The overall structure of the DGCA V-Net model is shown in *Figure 2*. This model adopts the basic encoder-decoder architecture of three-dimensional V-Net to perform two operations of extracting and restoring lung nodule image features, respectively. In contrast to U-Net, V-Net uses a 5×5×5 convolution operation to expand the receptive field, while the downsampling operation is completed by a 2×2×2 convolution kernel with a stride of 2. Using convolution operations to replace pooling operations will cause the network to occupy less memory during training. Moreover, 1×1×1 point convolution is used to complete the residual connection (25) between the input of the first layer of convolution and the output of the last layer of convolution in the module, such that the original features can be retained. This renders the learning of the network smoother and more stable and further improves the accuracy and generalization ability of the model. During training, problems of gradient disappearance and gradient

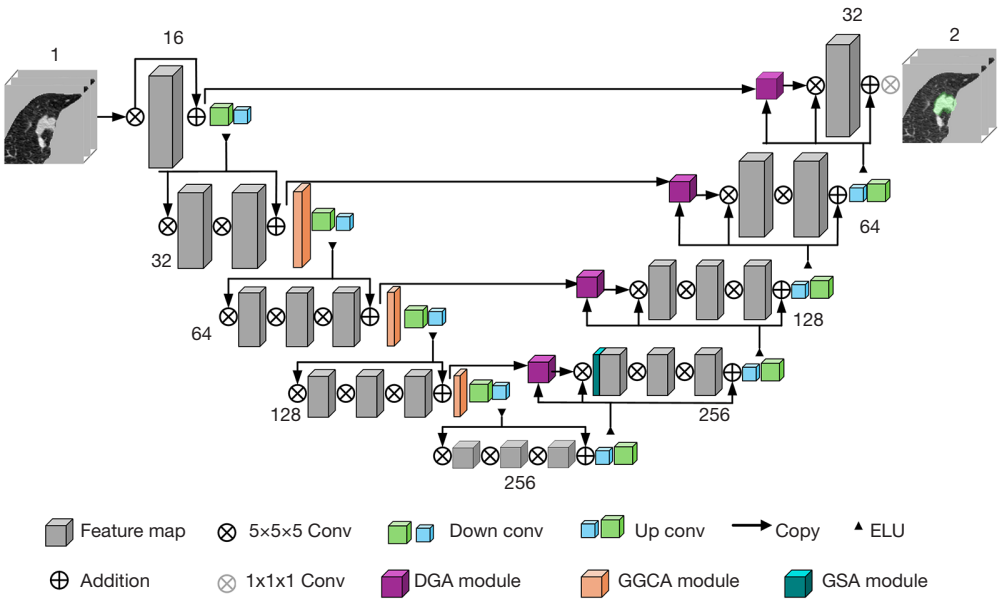

Figure 2 The basic structure of DGCA V-Net. The changes in the number of image channels caused by some operations in the figure have been marked above these operations. DGCA V-Net is an end-to-end network that integrates the GGCA module, the GSA module, and the DGA module into a V-Net model. Conv, convolution; DGA, dual-input-guided feature aggregation; DGCA, dual-channel grouped cross-dimension attention; ELU, exponential linear unit; GGCA, global grouped coordinate attention; GSA, grouped split attention.

explosion can be avoided and network convergence can be accelerated. To this base model, the GGCA module, GSA module, and DGA module are added. The specific operations are as follows:

First, in the encoding path, a new GGCA module is introduced after each layer of the original 1 to 3 repetitions of 5×5×5 3D convolution. This module can effectively capture the correlation between nodules and improve the expressive ability of the model.

Second, in the decoding path, the GSA module is introduced when deep features are restored to original features. This module effectively reorganizes and aggregates features of different branches through grouping and finer-grained splitting so as to better capture more detailed features and improve the segmentation result.

Finally, in order to reduce the loss of a large amount of detailed information during decoding, the DGA module is introduced in the skip connection. This module reduces the redundancy problem caused by high and low resolutions, enhances the learning of important features, and further improves the model's detailed segmentation of nodules.

The DGCA V-Net receives as input a single-channel grayscale image of pulmonary nodules. Therefore, the input size of the model is 1×16×96×96, corresponding to the number of channels, depth, height, and width, respectively. The output of the module includes nodule areas and non-nodule areas. Therefore, the output size is 2×16×96×96. The following sections include an in-depth analysis and detailed explanations of the GGCA module, GSA module, and DGA module.

GGCA module

To address V-Net's limitations in the down sampling process—specifically, its insufficient feature representation capacity and inability to effectively capture multidimensional global information, which leads to poor contextual understanding when processing complex nodules—we applied CA and devised a novel GGCA module. There are fewer parameters in GGCA than in the current CA mechanism. Its primary benefit is the creative incorporation of global feature data into the 3D spatial dimension. The depiction of key characteristics is improved by GGCA, which creates multidirectional attention maps by combining global information along the depth, height, and breadth axes. Notably, GGCA integrates positional information into channel attention, which allows the model to capture feature dependencies and long-range interactions

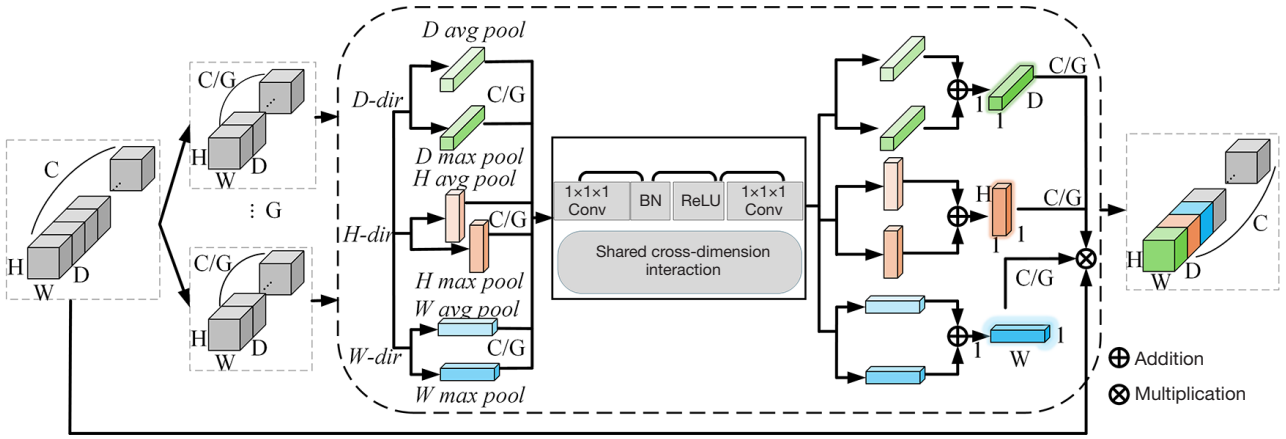

Figure 3 The basic structure of the GGCA module. Avg pool, average pooling; BN, batch normalization; C, channels; Conv, convolution; D, depth; dir, direction; G, group; GGCA, global grouped coordinate attention; H, height; ReLU, rectified linear unit; W, width.

and better grasp interchannel linkages. To fully extract the multidimensional global context, efficiently minimize information loss, and greatly enhance the completeness of feature representation, the module combines the strategies of global average pooling and max pooling along various directions. *Figure 3* depicts the overall architecture of the module.

The module is divided into five steps: (I) feature grouping, (II) global pooling, (III) shared convolutional layer, (IV) attention weight calculation, and (V) application of attention weights.

(I) In feature grouping, first, input feature $\mathbf{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$ is divided into G groups according to the number of channels, with each group containing C/G channels. Here, B is the batch size; C is the number of channels; and D, H, and W are the depth, height, and width of the feature map, respectively. The grouped feature map is represented as $\mathbf{X} \in \mathbb{R}^{B\times G\times \frac{C}{G}\times D\times H\times W}$. In contrast to the conventional CA module, the GGCA module minimizes the computational load for each group by grouping the input feature maps along the channel dimension using a grouped processing technique. By using a multigroup parallel processing approach, GGCA ensures that feature information from various channel groups can be separately extracted and used, preserving the richness and diversity of feature representations.

(II) In global pooling, for the grouped feature map, pooling kernels with dimensions (D, 1, 1), (1, H, 1), and (1, 1, W) are applied with global average pooling and global max pooling operations in the three directions of depth, height, and width, respectively, to generate multiple perceptual feature maps. The depth expression of the *C/G-th* channel with depth D after global average pooling and global max pooling is shown in Eqs. [1,2] below:

$$\mathbf{X}_{d,avg} = \mathrm{AvgPool}(\mathbf{X}) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times D\times 1\times 1} \quad [1]$$

$$\mathbf{X}_{d,max} = \mathrm{MaxPool}(\mathbf{X}) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times D\times 1\times 1} \quad [2]$$

The height expression of the *C/G-th* channel with height H after global average pooling and global max pooling is shown in Eqs. [3,4] below:

$$\mathbf{X}_{h,avg} = \mathrm{AvgPool}(\mathbf{X}) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times 1\times H\times 1} \quad [3]$$

$$\mathbf{X}_{h,max} = \mathrm{MaxPool}(\mathbf{X}) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times 1\times H\times 1} \quad [4]$$

The width expression of the *C/G-th* channel with width W after global average pooling and global max pooling is shown in Eqs. [5,6] below:

$$\mathbf{X}_{w,avg} = \mathrm{AvgPool}(\mathbf{X}) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times 1\times 1\times W} \quad [5]$$

$$\mathbf{X}_{w,max} = \mathrm{MaxPool}(\mathbf{X}) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times 1\times 1\times W} \quad [6]$$

Through global average pooling and max pooling in multiple directions of depth, height, and width, respectively, multidimensional global information is captured, and the comprehensiveness

of feature extraction is improved.

(III) In the shared convolutional layer step, a shared convolutional layer is applied for feature processing for each grouped feature map. The shared convolutional layer consists of two 1×1×1 convolutional layers, a batch normalization (BN) layer, and a rectified linear unit (ReLU) activation function, which is used to reduce and restore the channel dimension. The depth expression output after pooling in the depth direction and the passage through the shared convolutional layer is shown in Eq. [7] below:

$$\mathbf{Y}_{\mathrm{d,avg}} = \mathrm{Conv}\left(\mathbf{X}_{\mathrm{d,avg}}\right), \mathbf{Y}_{\mathrm{d,max}} = \mathrm{Conv}\left(\mathbf{X}_{\mathrm{d,max}}\right) \quad [7]$$

The height expression output after pooling in the height direction and passage through the shared convolutional layer is shown in Eq. [8] below:

$$\mathbf{Y}_{\mathrm{h,avg}} = \mathrm{Conv}\left(\mathbf{X}_{\mathrm{h,avg}}\right), \mathbf{Y}_{\mathrm{h,max}} = \mathrm{Conv}\left(\mathbf{X}_{\mathrm{h,max}}\right) \quad [8]$$

The width expression output after pooling in the width direction and passage through the shared convolutional layer is shown in Eq. [9] below:

$$\mathbf{Y}_{\mathrm{w,avg}} = \mathrm{Conv}\left(\mathbf{X}_{\mathrm{w,avg}}\right), \mathbf{Y}_{\mathrm{w,max}} = \mathrm{Conv}\left(\mathbf{X}_{\mathrm{w,max}}\right) \quad [9]$$

The importance of features varies across different channels. Ordinary convolution operations cannot dynamically adjust the importance of features and cannot highlight key features. Therefore, through the shared convolutional layer and attention mechanism, attention maps in the depth, height, and width directions are generated to weight the input feature map, enhance the expression of important features, and suppress unimportant features.

(IV) In the attention weight calculation, through the addition of the outputs of the convolutional layers and application of the sigmoid activation function, the attention weights in the depth, height, and width directions are generated, as shown in Eqs. [10-12] below:

$$\mathbf{A}_{\mathrm{d}} = \sigma\left(\mathbf{Y}_{\mathrm{d,avg}} + \mathbf{Y}_{\mathrm{d,max}}\right) \in \mathbb{R}^{\mathrm{B}\times\mathrm{G}\times\frac{\mathrm{C}}{\mathrm{G}}\times\mathrm{D}\times1\times1} \quad [10]$$

$$\mathbf{A}_{\mathrm{h}} = \sigma\left(\mathbf{Y}_{\mathrm{h,avg}} + \mathbf{Y}_{\mathrm{h,max}}\right) \in \mathbb{R}^{\mathrm{B}\times\mathrm{G}\times\frac{\mathrm{C}}{\mathrm{G}}\times1\times\mathrm{H}\times1} \quad [11]$$

$$\mathbf{A}_{\mathrm{w}} = \sigma\left(\mathbf{Y}_{\mathrm{w,avg}} + \mathbf{Y}_{\mathrm{w,max}}\right) \in \mathbb{R}^{\mathrm{B}\times\mathrm{G}\times\frac{\mathrm{C}}{\mathrm{G}}\times1\times1\times\mathrm{W}} \quad [12]$$

where σ represents the sigmoid activation function.

(V) In the application of attention weights, the input feature map is finally weighted according to the attention weights to obtain the output feature map, as shown in Eq. [13] below:

$$\mathbf{Z} = \mathbf{X} \times \mathbf{A}_{\mathrm{d}} \times \mathbf{A}_{\mathrm{h}} \times \mathbf{A}_{\mathrm{w}} \in \mathbb{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{D}\times\mathrm{H}\times\mathrm{W}} \quad [13]$$

Here, the attention weights $\mathbf{A}_{\mathrm{d}}$, $\mathbf{A}_{\mathrm{h}}$, and $\mathbf{A}_{\mathrm{w}}$ can be expanded in the depth, height, and width directions, respectively, to match the size of the input feature map. The GGCA model combines multidimensional global information and an attention mechanism. Through the capture of multidimensional global information and feature enhancement, the performance of the model is significantly improved.

GSA module

Some details may be lost because the decoder's small receptive field size makes it impossible to successfully restore the deep feature map to the original feature map. To address this, we apply the GSA module, drawing inspiration from the dynamic convolutional kernel selection method of the SK-Net and the channel attention mechanism of the SE-Net. By combining the benefits of both, GSA constitutes a two-channel feature-processing mechanism in contrast to SE-Net and SK-Net. To accomplish the hierarchical processing of channel information, the feature map is separated into several subgroups in the channel dimension. Each subgroup is simultaneously subjected to a fine-grained feature splitting procedure. A feature aggregation unit with distinct semantics is created by adaptively merging the divided feature representations using weighted aggregation. This architecture achieves collaborative modeling of spatial context and channel dependency by improving the network's capacity to capture multiscale features and efficiently extract cross-group information. *Figure 4* presents the GSA module's fundamental structural diagram.

As can be seen from *Figure 4*, the size of the input feature map is $C \times D \times H \times W$. The module is divided into three parts. The first part involves a 3×3×3 three-dimensional grouped convolution, BN (26), and ReLU activation function on the input feature map. The input feature map is first divided into groups, with the number of groups denoted by "g" (set to 2 in this experiment). Each group is then further split into "*r*" branches, where *r* denotes

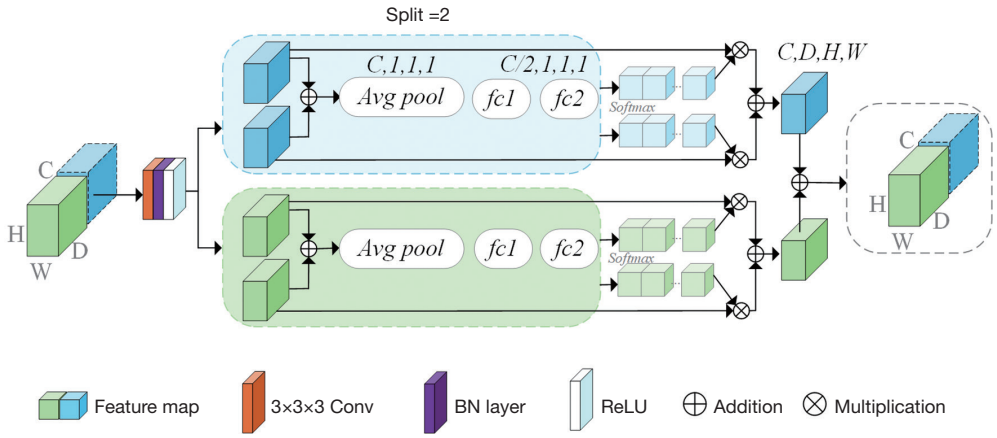

Figure 4 The basic structure of the GSA module. Avg pool, average pooling; BN, batch normalization; C, channels; D, depth; GSA, grouped split attention; H, height; ReLU, rectified linear unit; W, width.

the number of branches per group. Accordingly, the total number of branches of the obtained feature map can be expressed as $G = g \times r$, which equals 4 in this experiment. In addition, r is also used to control the number of output channels of the 3D grouped convolution, such that $c_{out} = c_{\text{int}} \times r$. At this point, the size of the feature map is $2C \times D \times H \times W$, which allows more branches to be placed within each group for feature extraction, which can improve the expressive ability of the model. The feature map is then reorganized to obtain a shape of $(B, r, rC/r, D, H, W)$. In this way, each branch has an independent channel. Subsequently, all branch features are aggregated to obtain a shape of (B, C, D, H, W). The purpose of this is to fuse the information of multiple branches together to obtain a richer feature. The reorganization and aggregation operations, through summing up the features, can effectively fuse information of multiple branches, enhance the robustness of the model, and reduce the possibility of overfitting.

In the second part, the channel attention mechanism is applied to each group. First, global average pooling is performed on the feature map. The size of the feature map is $C \times 1 \times 1 \times 1$. Dimension reduction is then applied through the first fully connected layer. The number of channels is calculated by the scaling factor m. The obtained number of channels is $C \times r \times m$. In the experiment for this study, the scaling factor was set to 0.25, with the size of the feature map thus being C/2×1×1×1. The scaling factor reduces the number of parameters in the network as compared to the SK-Net model, thereby reducing the complexity of the model and the risk of overfitting. BN and ReLU activation function operations are further performed. Finally, dimension increase is enacted via the second fully connected layer. The size of the feature map is $2C \times D \times H \times W$.

The third part involves the allocation of attention weights. Softmax normalization is applied to the output of the second fully connected layer, and the shape is adjusted to generate attention weights, enabling the model to focus on important features. Following this, weighted summation is performed on the features according to the attention weights. The purpose of doing this is to dynamically adjust the degree of attention to different features and enhance the feature representation ability. Finally, the output feature maps of each group are added element-wise to obtain the final output of the GSA module.

DGA module

Due to the issue of redundant information being produced when high-resolution and low-resolution features are joined at the V-Net skip link, learning may be positioned incorrectly, which could impact the nodule's detailed segmentation. Drawing inspiration from the AG, we add the DGA module to the V-Net skip connection structure. In contrast to the conventional AG module, the DGA module works to efficiently filter data during the feature fusion phase, greatly reducing the interference of redundant features and enhancing nodule localization and segmentation performance. By combining elements from various resolutions, this module adroitly maximizes the flow of information and makes it possible for crucial features to be prominently displayed. By boosting the model's sensitivity to minute variations and its comprehension of intricate structures, this feature aggregation technique

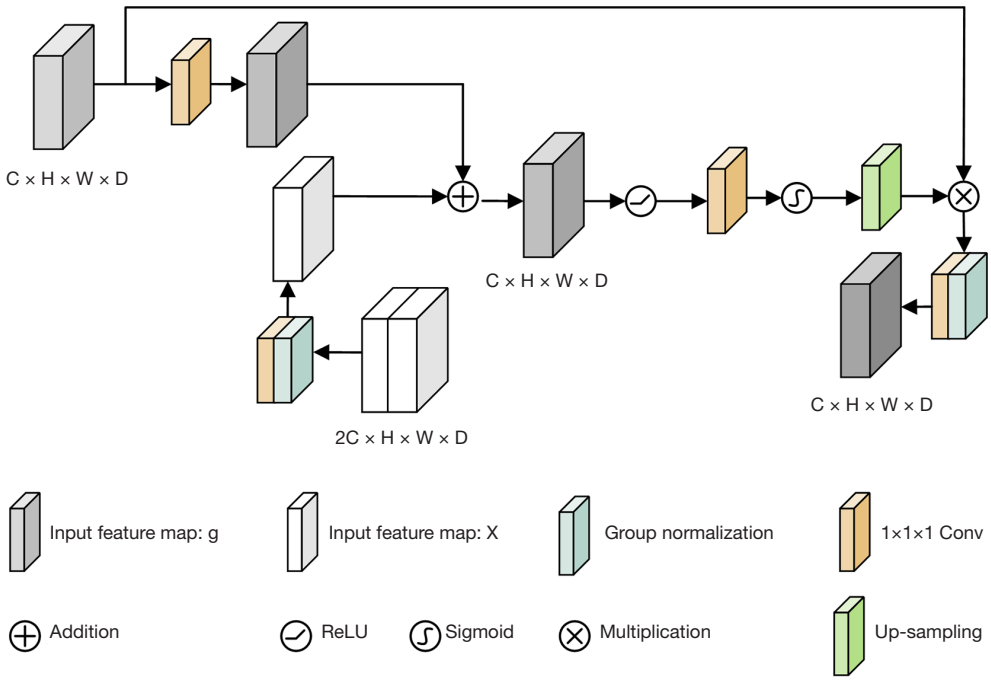

Figure 5 The basic structure of the DGA mechanism. C, channels; Conv, convolution; D, depth; DGA, dual-input-guided feature aggregation; H, height; ReLU, rectified linear unit; W, width.

strengthens the impact of detail segmentation and increases the model's stability and dependability when a variety of inputs are being handled. *Figure 5* illustrates the DGA module's fundamental architecture.

As can be seen from *Figure 5*, the DGA module has two input gates. One is the gated signal x of the low-resolution layer, and the other is the input signal g of the high-resolution layer. In the first step, 1×1×1 three-dimensional convolution and group normalization (27) are performed on signal *x* to adjust the number of channels from *2C* to *C* and to obtain the feature map $\mathbf{x}'$. Meanwhile, 1×1×1 three-dimensional convolution is performed on signal g to adjust the number of channels to *C* and obtain the feature map $\mathbf{g}'$. The size of both feature maps is $C \times H \times W \times D$. In the second step, element-wise addition of the feature maps $\mathbf{x}'$ and $\mathbf{g}'$ is completed, and then the activated feature map is obtained through the ReLU, as shown in Eq. [14] below:

$$\mathbf{u}_{x'+g'} = \sigma\left(\mathbf{W}_x^T\mathbf{x}' + \mathbf{W}_g^T\mathbf{g}' + \mathbf{b}_x + \mathbf{b}_g\right) \quad [14]$$

where σ is the ReLU activation function, $\mathbf{W}_x^T$ and $\mathbf{W}_g^T$ are convolution weights, and $\mathbf{b}_x$ and $\mathbf{b}_g$ are bias terms. In the third step, 1×1×1 three-dimensional convolution is performed to map the activated feature map to a lower-dimensional space for gating operations and thus reduce the training parameters. The sigmoid function is then executed to rescale each pixel value of the attention map to [0, 1]. The execution formula is expressed as follows:

$$\mathbf{a}_{x,g} = \delta\left(\mathbf{W}_\delta^T\left(\mathbf{C}^T\mathbf{u}_{x'+g'} + \mathbf{b}_c\right) + \mathbf{b}_\delta\right) \quad [15]$$

where δ is the sigmoid function, $\mathbf{W}_\delta^T$ and $\mathbf{C}^T$ are convolution weights, and $\mathbf{b}_c$ and $\mathbf{b}_\delta$ are bias terms. Upsampling is then performed, and the trilinear interpolation method is used to adjust the attention coefficient $\mathbf{a}_{x,g}$ to the same spatial size as the original feature map to achieve better context information fusion and obtain the processed attention coefficient $\mathbf{a}'_{x,g}$. In the final step, the element-wise multiplication of the originally input signal g and the attention coefficient $\mathbf{a}'_{x,g}$ is performed, and 1×1×1 three-dimensional convolution is completed to map the channels to the original size, and then normalization is completed to accelerate training convergence. The final attention output can be expressed as follows:

$$\mathbf{A}_{out} = \mathrm{GroupNorm}_{x,g}\left(\left(\mathbf{W}_{out}^T\left(\mathbf{a}'_{x,g} \times g\right)\right) + \mathbf{b}_{out}\right) \quad [16]$$

where $\mathrm{GroupNorm}_{x,g}$ is the group normalization operation,

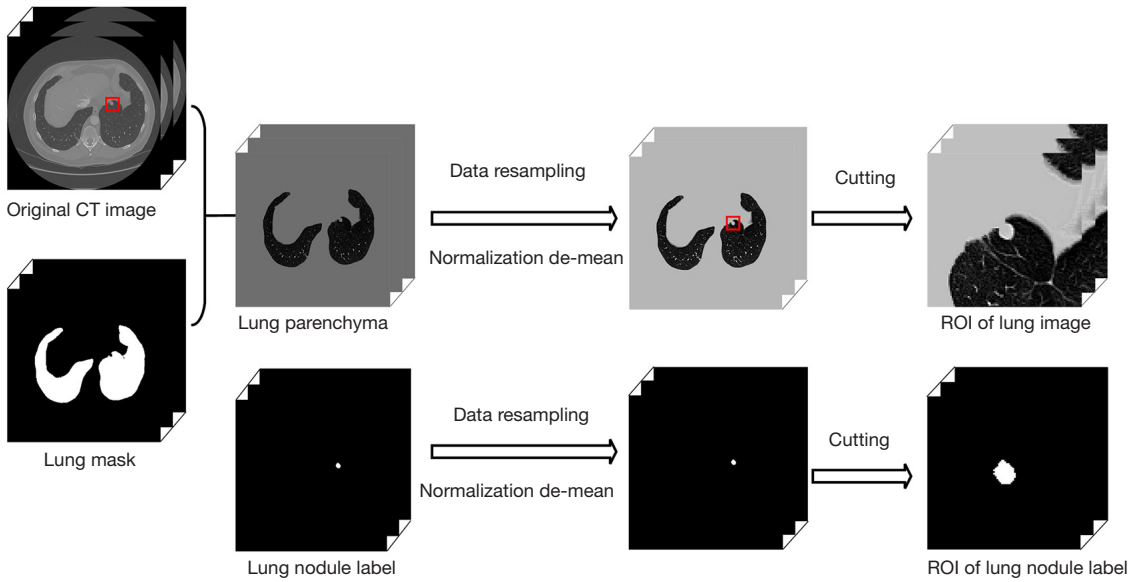

Figure 6 Image preprocessing flowchart. CT, computed tomography; ROI, region of interest.

$\mathbf{W}_{out}^{T}$ is the convolution weight, and $\mathbf{b}_{out}$ is the bias term.

Experiments

Datasets

In our evaluation, we used the publicly available Lung Nodule Analysis 2016 (LUNA16) dataset (28), which was derived from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset (29). The original LIDC-IDRI dataset contains 1,018 low-dose lung CT scans. By excluding scans with a slice thickness greater than 3 mm, inconsistent slice spacing, or missing slices, the LUNA16 dataset provides a curated subset of 888 CT scans with 1,186 annotated pulmonary nodules. The study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments.

Data preprocessing

The LIDC-XML-only files in LUNA16 and LIDC-IDRI were preprocessed to produce original photos, label images, and nodule information files. The LIDC XML-only files include the location and diameter annotation information recorded by clinicians for 1,186 pulmonary nodules. The data in the LUNA16 dataset may have different formats, resolutions, sizes, etc., and these factors may affect the segmentation performance of the model. Therefore, it was necessary to preprocess the data. As shown in *Figure 6*, preprocessing included lung area mask superposition, lung window interception, unified pixel spacing, mean normalization, and image cropping.

In preprocessing, the first step was lung area mask superposition. Since the lung area mask values provided by the dataset were 3 for the left lung and 4 for the right lung, it was necessary to first unify the lung area mask values to 1 and then multiply this value by that of the original CT image. The second step was to unify the pixel spacing of the original CT image and label to 1 mm. The third step included normalization and mean removal. Since lung tissue is relatively complex and there are a large number of lung trachea, lung blood vessels, tissue mucosa, and other structures around pulmonary nodules. Therefore, before normalization and mean removal, the gray value in the CT image was converted to Hounsfield unit (HU) values. The range was intercepted close to a lung density of –1,000 to 400, where values greater than 400 HU were set to 400 and values lower than –1,000 HU were set to –1,000. Normalization was used by subtracting the minimum value of the lung window (–1,000) from the original value and dividing by the lung width (the maximum value of the lung window minus the minimum value of the lung window). Finally, mean removal was performed. The mean value in the LUNA16 competition is approximately 0.25, and thus this value was subtracted from all numbers after

normalization. In the fourth step, the processed image was cropped and labeled into a volume block of 16×96×96, where 16 is the number of image layers, and 96×96 is the size of each layer of the image. The dataset was divided into training and test sets at a ratio of 9:1 according to the number of cases to avoid data leakage and ensure a fair evaluation (13).

Evaluation metrics

The metrics employed in the evaluation of this model consisted of the Dice similarity coefficient (DSC), intersection over union (IoU), precision, and recall.

DSC is an indicator used to evaluate the similarity between labeled samples and model-segmented samples. Its value range is between 0 and 1. The closer the value is to 1, the more similar the segmentation result. It can be expressed by the following confusion matrix:

$$DSC = \frac{2TP}{FP + 2TP + FN} \quad [17]$$

Where TP (true positive) represents the number of positive samples that were correctly identified as positive, FN (false negative) represents the number of positive samples that were incorrectly classified as negative, and FP (false positive) represents the number of negative samples that were incorrectly classified as positive.

IoU is a commonly used standard evaluation index in segmentation problems and represents the true pulmonary nodule area and the pulmonary nodule area obtained by model segmentation. It is usually used to measure the degree of overlap between the predicted box and the true box. It can be expressed by the following confusion matrix:

$$IoU = \frac{TP}{TP + FP + FN} \quad [18]$$

Precision, a metric used to assess the accuracy of the model's prediction outputs, is the percentage of true positives among all of the model's positive predictions. Recall, on the other hand, measures the percentage of true positive classes that are successfully identified, which indicates the model's capacity to find positive class targets across the whole dataset. These two indicators are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \quad [19]$$

$$Recall = \frac{TP}{TP + FN} \quad [20]$$

Parameter setting

The LUNA16 dataset was used in this work for model training, with 90% of the dataset being used for network model training and 10% for testing. In the experimental stage, the initial learning rate was set to 0.001, and a periodic function was introduced to prompt the learning rate to gradually change in a cosine-like pattern. This approach helped to avoid the model falling into local minima and thus to better converge to the optimal solution. With stochastic gradient descent (SGD) selected as the optimizer, the initial momentum was set to 0.99 and decreased to 0.9 in the 180th round, while the weight decay was set to 1E–8. In addition, the value of epoch was 200, while the batch size was 4. Meanwhile, we compared three optimizers: Adaptive Moment Estimation (Adam), Adam with decoupled weight decay (AdamW), and SGD—all of which were implemented under the same network structure and training setup.

The hardware environment used in this experiment included an A100-PCIE graphics card (NVIDIA, Santa Clara, CA, USA) with a graphics card memory of 40 GB. The model was implemented with Python 3.11.5 (Python Software Foundation, Wilmington, DE, USA) and the PyTorch2.1.1 + CUDA12.4 deep learning framework.

Results

To evaluate the performance of the DGCA V-Net model developed in this study in the lung nodule segmentation task, it was compared with several existing representative models, including those with transformer architecture (e.g., WingsNet, SegNet, and TransUNet) and CNN structure (e.g., UNet++, nnU-Net, and scSE-NL V-Net). The evaluation metrics included DSC, IoU, precision, and recall, with the experimental results being presented in *Table 1*.

Multiple indicators revealed notable advantages of DGCA V-Net when compared to the baseline model V-Net: the DSC increased from 0.7564 to 0.7921 (a 4.72% increase), the IoU increased from 0.6231 to 0.6662, the precision increased from 0.7874 to 0.8102, and the recall increased from 0.7627 to 0.7993. This suggests that the model has improved stability and segmentation accuracy. In comparison to other models, DGCA V-Net obtained the highest DSC metric value (0.7921), outperforming CNN-based models ResDSda_U-Net (0.7791) and nnU-Net (0.7911) as well as the transformer-based models SegNet (0.7862) and TransUNet (0.7875). As for the precision

Table 1 Comparison of different pulmonary nodule segmentation models

Type	Model	DSC	IoU	Precision	Recall
Transformer	WingsNet (30,31)	0.7919	0.6790	0.7858	0.8554†
	MS-TCNet (31,32)	0.7720	0.6581	0.7712	0.8269
	SegNet (31,33)	0.7862	0.6766	0.7749	0.8521
	CIPA (Mamba) (34)	–	0.6381	–	–
	TransUNet (35)	0.7875	0.6629	0.8079	0.7866
CNN	ResUnet (36)	0.7619	0.6302	0.8091	0.7592
	MS-UNet (37)	0.7740	–	–	0.8260
	SMR-UNet (38)	0.7785	0.6541	0.7982	0.7970
	3D-UNet-CRF (39)	0.7830	–	–	–
	QAU-Net (40)	0.7744	–	–	–
	MA-Net (41)	0.7490	–	–	–
	MSDS-UNet (42)	0.7680	–	–	0.7870
	DC-UNet (43)	0.7703	–	–	–
	MC-DFN (44)	0.7486	0.6330	0.7586	0.8044
	nnU-Net (45)	0.7911	0.6650	0.8021	0.8032
	UNet++ (46)	–	0.7705†	–	–
	ResDSda_U-Net (15)	0.7791	0.6511	0.8084	0.7876
	SCA-VNet (13)	0.7708	0.6409	0.7924	0.7818
	scSE-NL V-Net (14)	0.7604	0.6272	0.8010	0.7669
	V-Net (12)	0.7564	0.6231	0.7874	0.7627
	DGCA V-Net (ours)	0.7921†	0.6662	0.8102†	0.7993

†, the optimal result. CIPA, cross-modal interactive perception network with Mamba; CNN, convolutional neural network; DC-UNet, dilated convolution U-Net; DGCA V-Net, dual-channel grouped cross-dimension attention V-Net; DSC, Dice similarity coefficient; IoU, intersection over union; MA-Net, multiscale attention net; MC-DFN, multiscale cross-domain deep fusion network; MS-TCNet, multiscale transformer convolutional neural network; MS-UNet, multiscale U-Net; MSDS-UNet, multiscale deeply supervised 3D U-Net; QAU-Net, quartet attention U-Net; ResDSda_U-Net, residual dense spatial and dual attention U-Net; SCA-VNet, Sobel-enhanced 3D coordinate attention and atrous spatial pyramid pooling V-Net; scSE-NL V-Net, spatial and channel squeeze-and-excitation with nonlocal V-Net; SMR-UNet, self-attention and multiscale feature residual U-Net; TransUNet, transformer U-Net; 3D-UNet-CRF, 3D U-Net combined with 3D conditional random field optimization.

metric, DGCA V-Net outperformed all of the compared models, including WingsNet (0.7858). Although WingsNet (0.8554) and SegNet (0.8521) had somewhat higher recall scores, DGCA V-Net nevertheless achieved a comparatively high score of 0.7993, surpassing most CNN models such as scSE-NL V-Net (0.7669). In terms of the IoU metric, the DGCA V-Net yielded a score of 0.6662. Although this was slightly lower than that of UNet++ (0.7705), it outperformed most CNN models, such as SCA-VNet (0.6409). In summary, the DGCA V-Net outperformed the majority of representative models in the critical metrics of DSC and precision in the experimental setup we established. Although it underperformed in some indicators, for instance, demonstrating an inferior IoU as compared to UNet++, its overall performance was stable, indicating that the proposed structure has the potential to increase the precision and resilience of pulmonary nodule segmentation.

Furthermore, we conducted a visual analysis of the segmentation results obtained from the proposed DGCA V-Net model on various types of pulmonary nodules. In

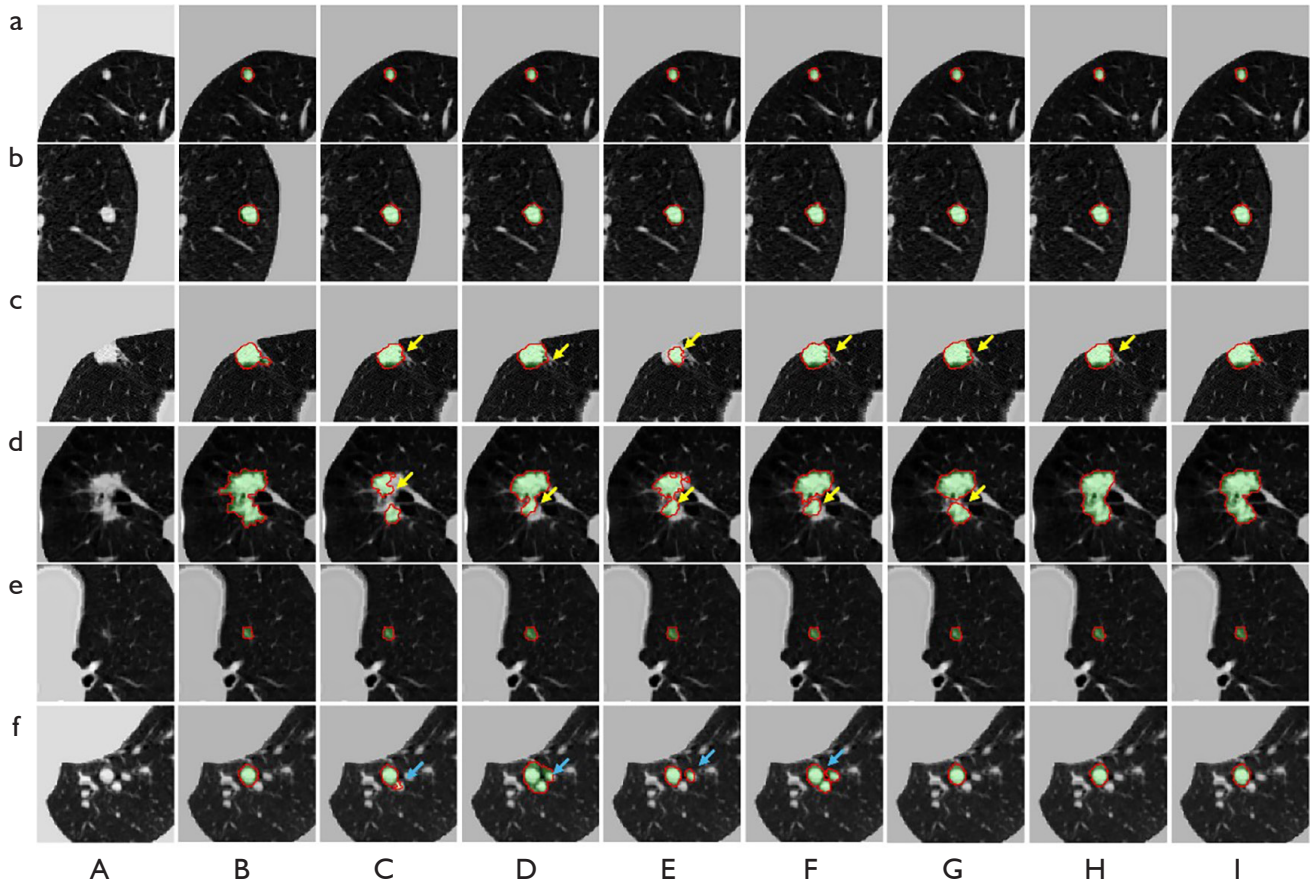

Figure 7 Segmentation results of DGCA V-Net and comparative models. The blue arrows indicate the oversegmented regions of each model, while the yellow arrows indicate the regions where the segmentation effect of each model is poor. (a: small nodules; b: medium nodules; c: large nodules; d: adherent nodules; e: ground-glass opacity nodules; f: calcified nodules). (A) Image, (B) ground truth, (C) V-Net, (D) SCA-VNet, (E) scSE-NL-VNet, (F) ResDSda_U-Net, (G) TransUNet, (H) nnUNet, (I) DGCA V-Net. DGCA V-Net, dual-channel grouped cross-dimension attention V-Net; ResDSda_U-Net, residual dense spatial and dual attention U-Net; SCA-VNet, Sobel-enhanced 3D coordinate attention and atrous spatial pyramid pooling V-Net; scSE-NL-VNet, spatial and channel squeeze and excitation with nonlocal V-Net; TransUNet, Transformer U-Net.

Figure 7, each row in the figure corresponds to a specific type of nodule, including (a) small nodules, (b) medium-sized nodules, (c) large nodules, (d) adherent nodules, (e) ground-glass nodules, and (f) calcified nodules. Each column sequentially presents the segmentation results for (A) image, (B) ground truth, (C) V-Net, (D) SCA-VNet, (E) scSE-NL V-Net, (F) ResDSda U-Net, (G) TransUNet, (H) nnUNet, and (I) DGCA V-Net. It is evident from *Figure* 7 that for small nodules, medium-sized nodules, and ground-glass opacities, most models achieve relatively accurate segmentation, demonstrating good sensitivity and localization capabilities for the lesion area. Conversely, for nodule types characterized by complex boundary structures or significant morphological variations, such as large nodules and adherent nodules, the performance of traditional models is markedly inadequate.

Figure 7 also shows that in large nodules (row c), V-Net, SCA-VNet, scSE-NL V-Net, ResDSda U-Net, TransUNet, and nnUNet demonstrated varying degrees of undersegmentation, as indicated by the yellow arrows (columns 4–8). These models exhibited limited capabilities in modeling scale variations and spatial distributions during feature extraction, which resulted in an inability to fully delineate the entire area of the lesions. This issue was more pronounced in adherent nodules (row d), particularly for SCA-VNet, scSE-NL V-Net, ResDSda

Table 2 Comparison of different optimizers on the segmentation performance of the proposed model on the LUNA16 dataset

Optimizer	DSC	IoU	Precision	Recall
Adam (47)	0.7663	0.6401	0.7882	0.7775
AdamW (48)	0.7770	0.6518	0.8255†	0.7648
SGD (ours)	0.7921†	0.6662†	0.8102	0.7993†

†, the optimal result. Adam, adaptive moment estimation; AdamW, Adam with decoupled weight decay; DSC, Dice similarity coefficient; IoU, intersection over union; LUNA16, Lung Nodule Analysis 2016; SGD, stochastic gradient descent.

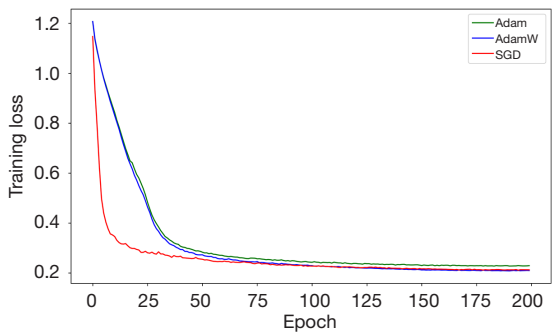

Figure 8 Comparison of training losses across different optimizers on the LUNA16 dataset. Adam, adaptive moment estimation; AdamW, Adam with decoupled weight decay; LUNA16, Lung Nodule Analysis 2016; SGD, stochastic gradient descent.

U-Net, and TransUNet, which demonstrated significant undersegmentation (yellow arrows in columns 4–7). Furthermore, for calcified nodules (row f), some models, such as SCA-VNet, scSE-NL V-Net, and ResDSda U-Net, exhibited oversegmentation (blue arrows in columns 4–6), misidentifying surrounding nonlesion tissues as part of the nodule. This suggests that these models exhibit an excessive response to feature expression when addressing areas with complex adjacent tissues, resulting in an ineffective balance between local and global information.

In contrast, DGCA V-Net more effectively aligned with the true labels across various nodule types, significantly minimizing both oversegmentation and undersegmentation issues. Notably, it exhibited enhanced robustness and generalization capabilities in the detection of adherent and calcified nodules.

Discussion

To assess the effectiveness of the proposed DGCA V-Net model in enhancing the accuracy of lung nodule segmentation, we conducted ablation experiments using the LUNA16 dataset. We used V-Net as the baseline model and performed a comprehensive comparison and analysis between it and DGCA V-Net to evaluate the effectiveness of each module. We compared and analyzed the segmentation performance, computational efficiency (including training and testing time), impact of various attention mechanisms on performance, and the number of parameters.

Ablation experiments on the optimizer

Regarding the choice of optimization algorithms, we used the LUNA16 dataset to compare three optimizers: AdamW, Adam, and SGD—all of which were implemented under the same network structure and training setup. *Table 2* displays the findings of these comparisons. The SGD optimizer delivered the best segmentation performance and performed well on the majority of assessment criteria. In addition, we conducted systematic comparative experiments under various initial conditions for different optimizers, as well as multiple repeated experiments on three mainstream optimizers under multiple random seeds (Tables S1,S2).

Additionally, to investigate the performance of several optimizers during the training process, we plotted the curves of the training loss of three optimizers changing with the epoch under the same network structure and training settings. The findings in *Figure 8* show that the SGD optimizer converges faster during the early stage of training, with a lower overall loss value, indicating greater convergence stability and optimization efficiency. As a result, in the subsequent studies, we used SGD as the optimizer.

Ablation experiments

We conducted ablation tests using the LUNA16 dataset to confirm that the DGCA V-Net model is efficient in increasing the accuracy of pulmonary nodule segmentation. V-Net was used the foundational model, to which several

Table 3 Ablation experiment results of each module on the LUNA16 dataset

Model	DSC	IoU	Precision	Recall
V-Net	0.7564	0.6231	0.7874	0.7627
V-Net + GGCA	0.7725	0.6437	0.7577	0.8184
V-Net + GSA	0.7747	0.6426	0.7835	0.7946
V-Net + DGA	0.7842	0.6554	0.7973	0.7964
V-Net + GGCA + GSA	0.7691	0.6367	0.7568	0.8132
V-Net + GGCA + DGA	0.7815	0.6545	0.8043	0.7753
V-Net + GSA + DGA	0.7787	0.6507	0.7705	0.8220[†]
DGCA V-Net (ours)	0.7921[†]	0.6662[†]	0.8102[†]	0.7993

[†], the optimal result. DGA, dual-input-guided feature aggregation; DGCA V-Net, dual-channel grouped cross-dimension attention V-Net; DSC, Dice similarity coefficient; GGCA, global grouped coordinate attention; GSA, grouped split attention; IoU, intersection over union; LUNA16, Lung Nodule Analysis 2016.

modules were progressively added, and each module's impact on the model's performance was assessed. *Table 3* summarizes the outcomes of the ablation tests.

When complicated nodules are being examined, it is challenging to collect context information due to V-Net encoder's inadequate downsampled feature representation. To address this, we added the GGCA module to the encoder. Comparing the first and second rows of *Table 3*, we can see that the addition of GGCA improved the overall segmentation performance. The DSC increased by 2.12% to 0.7725, the IoU increased by 3.3% to 0.6437, and the recall increased by 7.30% to 0.8184, yet the precision only decreased by 3.91%. This suggests that the GGCA module can successfully improve the encoder's capacity to represent nodule correlations.

Secondly, we added the GSA module to the decoder's low-resolution layers to solve the issue of deep features having difficulty in recovering the original details due to the decoder's narrow receptive field. In *Table 3*, rows 1 and 3 show that following the implementation of GSA, the DSC increased by 2.41% to 0.7747, the IoU increased by 3.13% to 0.6426, the recall increased by 4.18% to 0.7946, and the precision only slightly decreased by 0.49%. This suggests that by using the channel grouping and weighting technique, the GSA module can successfully improve the feature restoration capability and lessen the loss of crucial information.

Furthermore, we added the DGA module to the skip connections to solve the issue in which the combination of high- and low-resolution features in the V-Net skip connections introduce redundant information. When the DGA module was added, the DSC increased by 3.67% to 0.7842, the IoU increased by 5.18% to 0.6554, the precision increased by 1.25% to 0.7973, and the recall increased by 4.41% to 0.7964, as shown in rows 1 and 4 of *Table 3*. This suggests that the segmentation performance can be improved by the DGA module's capacity to efficiently suppress duplicated features and improve the nodule region's expression ability.

Finally, we conducted a multimodule combination ablation experiment to assess the synergistic impact of combining the different modules. *Table 3* (rows 5–7) demonstrates how various module combinations increased performance across a range of evaluation metrics. For the vast majority of evaluation metrics, the final model (final row in *Table 3*) obtained the best segmentation performance, with a 4.72%, 6.92%, 2.90%, and 4.80% improvement over V-Net for the DSC, IoU, precision, and recall, and corresponding values of 0.7921, 0.6662, the 0.8102, and 0.7993. Overall, these findings suggest that the GGCA, GSA, and DGA modules may greatly increase the segmentation accuracy of pulmonary nodules and have good independence and combinability.

In order to thoroughly examine the distinct effects of each attention module on model performance, the segmentation visualization results of several model structures for six different types of lung nodules were evaluated with the LUNA16 public dataset (*Figure 9*). In complex settings such as those with small nodules, adherent nodules, and multiple nodules, the baseline model V-Net frequently exhibited issues such as undersegmentation (yellow arrow in *Figure 9*) and oversegmentation (blue arrow in *Figure 9*).

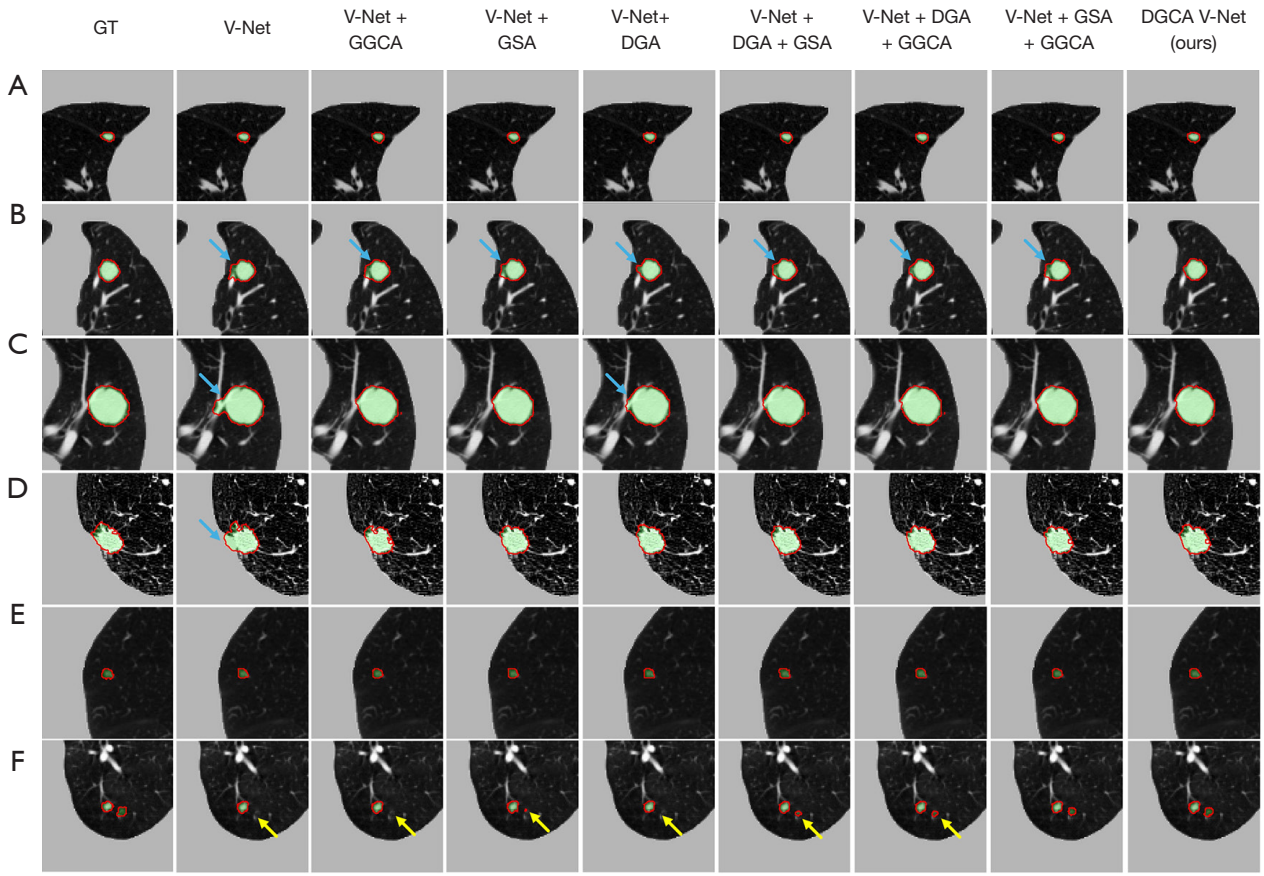

Figure 9 The visual results of the ablation study. The blue arrows indicate the oversegmented areas of each model, while the yellow arrows indicate the areas where the segmentation effect of each model is not satisfactory (A: small nodules; B: medium-sized nodules; C: large nodules; D: juxtapleural nodules; E: ground-glass opacity nodules; F: multiple nodules). Each column corresponds to the segmentation output of a different model configuration, arranged from left to right as follows: ground truth, the baseline V-Net, V-Net variants with individual attention modules (GGCA, GSA, and DGA), V-Net variants with combinations of two modules (DGA + GSA, DGA + GGCA, and GSA + GGCA), and finally the proposed complete model, DGCA V-Net. DGA, dual-input-guided feature aggregation; DGCA, dual-channel grouped cross-dimension attention; GGCA, global grouped coordinate attention; GSA, grouped split attention; GT, ground truth.

Columns 2 and 3 in *Figure 9* further show that the GGCA module considerably improved the model's segmentation quality in adherent and large nodules (rows C and D) and successfully mitigated the oversegmentation issue. Its modeling capacity for fine-grained local structures is constrained, however, as several nodules (row F in *Figure 9*) nonetheless exhibited a degree of undersegmentation. Moreover, in the comparison of column 2 and 4, it can be surmised that after introduction of the GSA module, the performance of the model in large nodules (row C) and adherent nodules (row D) also significantly improved, and the over-segmentation was significantly reduced; however, there was a certain degree of undersegmentation in multiple nodules (row F). Furthermore, by comparing columns 2 and 5 in *Figure 9*, we can see that after the DGA module was introduced, the model improved the boundary ability between nodules and the background in medium-sized nodules (row B) and adherent nodules (row D); however, slight oversegmentation occurred in large nodules (row C), and under-segmentation also occurred in multiple nodules (row F). Finally, the combination of multiple modules (columns 6–8 in *Figure 9*) further exerted their respective advantages, especially in the scenarios of multiple nodules (row F) and adherent nodules (row D), providing more

stable and excellent segmentation effects. The final proposed DGCA V-Net model integrates three attention mechanisms: GGCA, GSA, and DGA. It demonstrated a segmentation effect closest to the ground truth in all samples. Especially in complex structural scenarios (such as multiple nodules in row F and adherent nodules in row D), it demonstrated excellent robustness and accuracy.

Computational efficiency experiment

Apart from the improvement in segmentation accuracy, we also assessed DGCA V-Net's performance in terms of time efficiency, as shown in *Table 4*, in which the training and testing durations of the DGCA V-Net and baseline V-Net are contrasted. According to the results, V-Net and DGCA V-Net require similar amounts of time to train—11,862 and 11,892 seconds, respectively. This implies that training effectiveness is not significantly impacted by the attention module's inclusion. The testing period for DGCA V-Net was 64.23 seconds, which was marginally longer than V-Net's 58.94 seconds, but nonetheless acceptable given the markedly improved segmentation performance. Overall, DGCA V-Net provides an improvement in model convenience and representational capability without sacrificing efficiency.

Table 4 Comparison of model training and testing times

Model	Training time, s	Test time, s
V-Net	11,862	58.94
DGCA V-Net	11,892	64.23

DGCA V-Net, dual-channel grouped cross-dimension attention V-Net.

Attention mechanism ablation study

To evaluate the differences in performance between the attention module proposed in this paper and existing attention modules (such as CA, SE, and SK) in pulmonary nodule segmentation, we embedded multiple attention modules into the V-Net framework and conducted comparative experiments on the LUNA16 dataset. The results are presented in *Table 5*. The GGCA module applied in our study was compared against the CA module in the first set of studies (upper part of *Table 5*). V-Net + GGCA had a higher DSC (0.7725), IoU (0.6437), and recall (0.8184) than did V-Net + CA, which had values of 0.7694, 0.6386, and 0.7539, respectively. The performance of GSA was compared with traditional attention processes, including SE and SK, in the second set of studies (lower part of *Table 5*). The best results in terms of DSC (0.7747) and recall (0.7946) were obtained by V-Net + GSA. Although its IoU (0.6426) was lower than that of V-Net + SK (0.6446), the difference was extremely slight. Overall, the GSA and GGCA modules demonstrated exceptional performance across a number of metrics, confirming the efficacy of the attention mechanism.

In addition to the performance improvements, we also analyzed the differences in parameter overhead among the attention modules used (*Table 6*). To ensure a fair comparison, we counted the number of parameters for each module while keeping the number of input channels fixed at 16 (C=16). In the first group, we compared the proposed GGCA module with its design inspiration, the CA module. The results indicated that GGCA introduces only 0.03 K parameters, significantly reducing computational overhead as compared to the 0.40 K parameters of the CA module. In the second group, we compared the proposed GSA module with the classic attention mechanisms SE and

Table 5 Performance comparison of different attention mechanisms embedded in the V-Net Model for the segmentation task in the LUNA16 dataset

Model	DSC	IoU	Precision	Recall
V-Net + CA	0.7694	0.6386	0.8184	0.7539
V-Net + GGCA	0.7725	0.6437	0.7577	0.8184
V-Net + SE	0.7656	0.6331	0.8030	0.7529
V-Net + SK	0.7725	0.6446	0.8019	0.7690
V-Net + GSA	0.7747	0.6426	0.7835	0.7946

CA, coordinate attention; DSC, Dice similarity coefficient; GGCA, global grouped coordinate attention; GSA, grouped split attention; IoU, intersection over union; LUNA16, Lung Nodule Analysis 2016; SE, squeeze and excitation; SK, selective kernel.

Table 6 Parameter comparison of the proposed and existing attention modules (input channels =16)

Model	Parameter (K)
CA	0.40
GGCA	0.03
SE	0.03
SK	40.51
GSA	8.51

CA, coordinate attention; GGCA, global grouped coordinate attention; GSA, grouped split attention; SE, squeeze and excitation; SK, selective kernel.

SK, with the GSA module generating 8.51 K parameters. Although this was higher than the 0.03 K parameters of SE, it was substantially lower than the 40.51 K parameters of SK, demonstrating that GSA achieves a favorable balance between the model's expressive capability and parameter complexity.

Conclusions

In this study, we developed a novel pulmonary nodule segmentation approach, DGCA V-Net. First, the GGCA module is used in the encoding path to comprehensively capture multidimensional global information in multiple directions. This design enhances the model's ability to handle complex spatial relationships and improves the integrity of feature information. Subsequently, in the deep part of the decoding path, the GSA module conducts feature extraction through group splitting, effectively aggregating and reorganizing important features to better capture spatial and channel details. Finally, in addressing the issue of inaccurate positioning caused by the redundant information generated when high- and low-resolution features are combined at the skip connection, the feature information of the encoding path and the decoding path was used as the input and fed into the DGA module. This integration of features from different resolutions highlights important features and thus improves the effect of detailed segmentation. Our experiment made use of the LUNA16 dataset for training and testing, and the results indicated that this model performs outstandingly in the pulmonary nodule segmentation task.

Despite the encouraging performance of DGCA V-Net model in the public 3D medical image datasets, certain limitations should be noted. First, our algorithm encountered challenges in accurately segmenting certain complex nodule contours. Second, the model has not been validated on clinical datasets that include heterogeneous imaging protocols, scanner types, or patient demographic characteristics, all of which may significantly impact segmentation performance in practical applications. Additionally, the variations in data acquisition standards among hospitals and the limited computing resources of many clinical devices necessitate further investigation into the model's robustness and lightweight deployment strategies. Finally, while our design emphasizes 3D segmentation, its potential applicability in 2D biomedical image segmentation remains unexplored. Future work will assess the generalizability of these modules by integrating them into 2D frameworks such as MoNuSeg to evaluate their cross-dimensional versatility. These directions are crucial steps in translating our method into a practical clinical tool. Future research will concentrate on enhancing heterogeneous data adaptability, lightweight deployment, and cross-dimensional generality to improve the practical application value and universality of the model.

Acknowledgments

None.

Footnote

Funding: This work was supported by National Natural Science Foundation of China (Grant Nos. 62276092 and 62303167), Science and Technology Research of Henan Province (No. 252102210042), Doctor Fund Project of Henan Polytechnic University (No. B2022-11), and Key Scientific Research Project of Henan Province (Nos. 24A520017 and 25A520009).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-2434/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki

and its subsequent amendments.

References

1. Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, Li L, Wei W, He J. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent 2024;4:47-53.
2. Zhou Q, Fan Y, Wang Y, Qiao Y, Wang G, Huang Y, Wang X, Wu N, Zhang G, Zheng X, Bu H. China National Guideline of Classification, Diagnosis and Treatment for Lung Nodules (2016 Version). Chinese Journal of Lung Cancer 2016;19:793-8.
3. Abdel-Basset M, Chang V, Mohamed R. A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems. Neural Comput Applic 2021;33:10685-718.
4. Xu WB, Liu Y, Zhang HW. Research progress in image segmentation based on region growing. Beijing Biomedical Engineering 2017;36:317-22.
5. Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, et al. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. IEEE Trans Med Imaging 2021;40:3413-23.
6. Zunair H, Ben Hamza A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. Comput Biol Med 2021;136:104699.
7. Zunair H, Hamza AJA. Masked Supervised Learning for Semantic Segmentation. 2022. arXiv: 2210.00923.
8. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical image computing and computer-assisted intervention-MICCAI 2015. 2015:234-41.
9. Zhou Z, Gou F, Tan Y, Wu J. A Cascaded Multi-Stage Framework for Automatic Detection and Segmentation of Pulmonary Nodules in Developing Countries. IEEE J Biomed Health Inform 2022;26:5619-30.
10. Akila Agnes S, Arun Solomon A, Karthick K. Wavelet U-Net++ for accurate lung nodule segmentation in CT scans: Improving early detection and diagnosis of lung cancer. Biomedical Signal Processing and Control 2024;87:105509.
11. Zhang G, Yang Z, Jiang S. Automatic lung tumor segmentation from CT images using improved 3D densely connected UNet. Med Biol Eng Comput 2022;60:3311-23.
12. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA; 2016:565-71.
13. Liu J, Li Y, Li W, Li Z, Lan Y. Multiscale lung nodule segmentation based on 3D coordinate attention and edge enhancement. Electronic Research Archive 2024;32:3016-37.
14. Zhou J, Ye J, Liang Y, Zhao J, Wu Y, Luo S, Lai X, Wang J. scSE-NL V-Net: A Brain Tumor Automatic Segmentation Method Based on Spatial and Channel "Squeeze-and-Excitation" Network With Non-local Block. Front Neurosci 2022;16:916818.
15. Ji Z, Zhao Z, Zeng X, Wang J, Zhao L, Zhang X, Ganchev I. ResDSda_U-Net: A Novel U-Net-Based Residual Network for Segmentation of Pulmonary Nodules in Lung CT Images. IEEE Access 2023;11:87775-89.
16. Xu X, Du L, Yin D. Dual-branch feature fusion S3D V-Net network for lung nodules segmentation. J Appl Clin Med Phys 2024;25:e14331.
17. Wang Z, Men J, Zhang F. Improved V-Net lung nodule segmentation method based on selective kernel. Signal Image Video Process 2023;17:1763-74.
18. Dutande P, Baid U, Talbar S. LNCDS: A 2D-3D cascaded CNN approach for lung nodule classification, detection and segmentation. Biomed Signal Process Control 2021;67:102527.
19. Agnes SA, Anitha J. Efficient multiscale fully convolutional UNet model for segmentation of 3D lung nodule from CT image. J Med Imaging (Bellingham) 2022;9:052402.
20. Li D, Yuan S, Yao G. Pulmonary nodule segmentation based on REMU-Net. Phys Eng Sci Med 2022;45:995-1004.
21. Li X, Wang W, Hu X, Yang J. Selective Kernel Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA; 2020:510-9.
22. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2020;42:2011-23.
23. Hou Q, Zhou D, Feng J. Coordinate attention for efficient

mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021:13708-17.
24. Zeng Y, Tsui PH, Wu W, Zhou Z, Wu S. Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net. J Digit Imaging 2021;34:134-48.
25. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016:770-8.
26. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning; 2015:448-56.
27. Wu Y, He K. Group normalization. Proceedings of the European Conference on Computer Vision (ECCV); 2018:3-19.
28. Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard CVD, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1-13.
29. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011;38:915-31.
30. Zheng H, Qin Y, Gu Y, Xie F, Yang J, Sun J, Yang GZ. Alleviating Class-Wise Gradient Imbalance for Pulmonary Airway Segmentation. IEEE Trans Med Imaging 2021;40:2452-62.
31. Su HQ, Lei HJ, Liu ZY, Yao LS, Li SY, Lin H, Chen GL, Chen X, Lei BY. Feature fusion network for pulmonary nodule segmentation and EGFR classification using dual encoders. Expert Systems with Applications 2025;280:127523.
32. Ao Y, Shi W, Ji B, Miao Y, He W, Jiang Z. MS-TCNet: An effective Transformer-CNN combined network using multi-scale feature learning for 3D medical image segmentation. Comput Biol Med 2024;170:108057.
33. Kuang H, Wang Y, Liu J, Wang J, Cao Q, Hu B, Qiu W, Wang J. Hybrid CNN-Transformer Network With Circular Feature Interaction for Acute Ischemic Stroke Lesion Segmentation on Non-Contrast CT Scans. IEEE Trans Med Imaging 2024;43:2303-16.
34. Mei J, Lin C, Qiu Y, Wang Y, Zhang H, Wang Z, Dai D. Cross-Modal Interactive Perception Network with Mamba for Lung Tumor Segmentation in PET-CT Images. 2025. arXiv: 2503.17261.
35. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021. arXiv: 2102.04306.
36. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2016; 2016:424-32.
37. Li Z, Yang J, Xu Y, Zhang L, Dong W, Du B. Scale-aware Test-time Click Adaptation for Pulmonary Nodule and Mass Segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2023; 2023:681-91.
38. Hou J, Yan C, Li R, Huang Q, Fan X, Lin F. Lung Nodule Segmentation Algorithm With SMR-UNet. IEEE Access 2023;11:34319-31.
39. Wu W, Gao L, Duan H, Huang G, Ye X, Nie S. Segmentation of pulmonary nodules in CT images based on 3D-UNET combined with three-dimensional conditional random field optimization. Med Phys 2020;47:4054-63.
40. Hong L, Wang R, Lei T, Du X, Wan Y. Qau-Net: Quartet Attention U-Net for Liver and Liver-Tumor Segmentation. 2021 IEEE International Conference on Multimedia and Expo (ICME); 2021:1-6.
41. Fan T, Wang G, Li Y, Wang H. MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation. IEEE Access 2020;8:179656-65.
42. Yang J, Wu B, Li L, Cao P, Zaiane O. MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT. Comput Med Imaging Graph 2021;92:101957.
43. Chen KB, Xuan Y, Lin AJ, Guo SH. Lung computed tomography image segmentation based on U-Net network fused with dilated convolution. Comput Methods Programs Biomed 2021;207:106170.
44. Dong CX, Dai DW, Li ZF, Xu SH. A novel deep network with triangular-star spatial-spectral fusion encoding and entropy-aware double decoding for coronary artery segmentation. Information Fusion 2024;112:102561.
45. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, Wasserthal J, Koehler G, Norajitra T, Wirkert S, Maier-Hein KH. nnU-Net: Self-adapting Framework for

U-Net-Based Medical Image Segmentation. 2018. arXiv: 1809.10486.
46. Zhou ZW, Siddiquee MMR, Tajbakhsh N, Liang JM. UNet plus plus: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. Ieee Transactions on Medical Imaging 2020;39:1856-67.
47. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. 2014. arXiv: 1412.6980
48. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2017. arXiv: 1711.05101.

Cite this article as: Zhang L, Liu T, Liang Y, Li Y, Zhang W, Sun J. Dual-channel grouped cross-dimension attention V-Net for pulmonary nodule segmentation. Quant Imaging Med Surg 2025;15(10):8876-8896. doi: 10.21037/qims-24-2434