3DSN-Net: A 3-D Scale-Aware convNet With Nonlocal Context Guidance for Kidney and Tumor Segmentation From CT Volumes

Huisi Wu, *Senior Member, IEEE*, Baiming Zhang, Zhuoying Li, Jing Qin, *Senior Member, IEEE*, and Tong-Yee Lee, *Senior Member, IEEE*

***Abstract*—Automatic kidney and tumor segmentation from CT volumes is a critical prerequisite/tool for diagnosis and surgical treatment (such as partial nephrectomy). However, it remains a particularly challenging issue as kidneys and tumors often exhibit large-scale variations, irregular shapes, and blurring boundaries. We propose a novel 3-D network to comprehensively tackle these problems; we call it *3DSN-Net*. Compared with existing solutions, it has two compelling characteristics. First, with a new scale-aware feature extraction (SAFE) module, the proposed *3DSN-Net* is capable of adaptively selecting appropriate receptive fields according to the sizes of targets instead of indiscriminately enlarging them, which is particularly essential for improving the segmentation accuracy of the tumor with large scale variation. Second, we propose a novel yet efficient nonlocal context guidance (NCG) mechanism to capture global dependencies to tackle irregular shapes and blurring boundaries of kidneys and tumors. Instead of directly harnessing a 3-D NCG mechanism, which makes the number of parameters exponentially increase and hence the network difficult to be trained under limited training data, we develop a 2.5D NCG mechanism based on projections of feature cubes, which achieves a tradeoff between segmentation accuracy and network complexity. We extensively evaluate the proposed *3DSN-Net* on the famous KiTS dataset with many challenging kidney and tumor cases. Experimental results demonstrate our solution consistently outperforms state-of-the-art 3-D networks after being equipped with scale aware and NCG mechanisms, particularly for tumor segmentation.**

***Index Terms*—3-D convolutional neural networks (CNNs), kidney and tumor segmentation, nonlocal context guidance (NCG), scale-aware feature extraction (SAFE).**

I. Introduction

KIDNEY cancer is now the ninth most common form of cancer among men and the 14th most common form of cancer among women. There are estimated to be more than 400000 newly diagnosed cases of kidney cancer worldwide in 2018, according to the statistics conducted by the World Health Organization [1]. Despite the high-incidence rate of kidney tumors, patients are likely to be cured in the early stages. Therefore, early detection is essential for increasing the survival rate. Accurate kidney and tumor segmentation plays a significant role in early detection and screening. On the other hand, as clinical evidence has shown that partial nephrectomy has a similar effect to radical nephrectomy in treating relatively small tumors with a lower risk of cardiovascular events from the long-term perspective [2], more and more partial nephrectomy surgeries have been conducted in clinical practice. However, most of the cases are small tumors with about 15% metastasis rate, which demands a precise segmentation and diagnosis for subsequent surgery planning [3]. Nowadays, clinicians mainly rely on preoperative CT images to obtain kidney tumor morphology, volume, and other information to assess its complexity and aggressiveness. In this regard, precise segmentation and correct classification of tumors significantly affect the formulation of treatment plans as well as the effectiveness of treatment.

Manual segmentation of kidney and tumors from CT volumes by radiologists is time-consuming, tedious, and error-prone, resulting in a growing demand for automatic approaches in clinical practice. However, automatic kidney and tumor segmentation from CT volumes remains a challenging problem due to: 1) the diverse growth positions, sizes, and shapes of kidney tumors; 2) the accompanying metastasis of tumor cells; 3) the irregular shapes of kidney and tumor; and 4) the blurring boundaries between kidney and adjacent organs and between kidney and tumor (some challenging cases are shown in Fig. 1). Considerable attempts have been made to address the above problems to achieve accurate segmentation.

Manuscript received 7 December 2022; revised 11 April 2023 and 31 May 2023; accepted 25 June 2023. Date of publication 20 July 2023; date of current version 17 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61973221 and Grant 62273241; in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2019A1515011165; in part by the COVID-19 Prevention Project of Guangdong Province, China, under Grant 2020KZDZX1174; in part by the Major Project of the New Generation of Artificial Intelligence under Grant 2018AAA0102900; in part by the Hong Kong Research Grant Council under General Research Fund Scheme under Grant 15205919; in part by the Hong Kong RGC Theme-Based Research Scheme under Grant T45-401/22-N; in part by the Hong Kong RGC General Research Fund under Grant 15218521; and in part by the National Science and Technology Council, Taiwan, under Grant 111-2221-E-006-112-MY3. This article was recommended by Associate Editor Y. Zhou. *(Corresponding author: Huisi Wu.)*

Huisi Wu, Baiming Zhang, and Zhuoying Li are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: hswu@szu.edu.cn).

Jing Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong (e-mail: harry.qin@polyu.edu.hk).

Tong-Yee Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 70101, Taiwan (e-mail: tonylee@mail.ncku.edu.tw).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2023.3291369.

Digital Object Identifier 10.1109/TCYB.2023.3291369

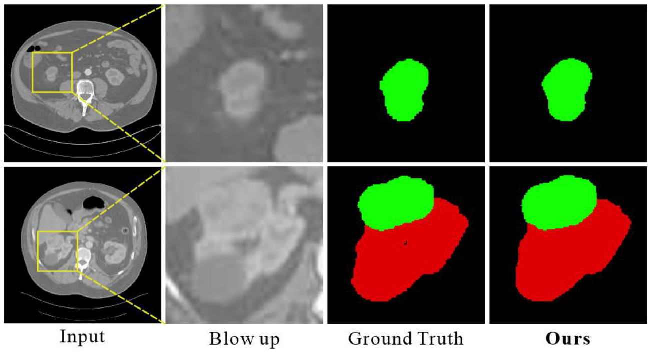


Fig. 1. Typical challenging cases for kidney and tumor segmentation, where tumors may appear with various scales, irregular shapes, and blurred boundaries. Red and green regions indicate kidney and tumors, respectively.

Traditional kidney and tumor segmentation approaches are often based on hand-crafted features, including intensities, textures, and shape priors [4]. For instance, Xie et al. [5] first employed Gabor filters to extract texture features and then used an expectation-maximization method to construct texture models to segment kidneys from CT images. Gloger et al. [6] proposed a multistep refinement approach to produce probability maps and then apply an extended level-set model to improve the segmentation results. Jin et al. [7] first used a 3-D generalized Hough transform (GHT) and a 3-D active appearance model (AAM) to complete the localization and then applied a modified random forest to yield the segmentation. However, these traditional methods often require manually adjusting parameters to obtain satisfactory performance and these parameters are usually very sensitive to noise and artifacts. More important, owing to limited representation capability, these hand-crafted features cannot achieve sufficient accuracy for clinical applications.

As deep learning develops, convolutional neural networks (CNNs) have been widely applied in various medical image segmentation tasks [8], [9]. Most existing medical image segmentation approaches are developed based on U-Net [10], which is an encoder–decoder architecture with multiple skip connections that can harness both spatial information and semantic information to improve the segmentation performance. For the 3-D image segmentation task, it is better to use a 3-D architecture rather than a 2-D network as the latter cannot take full advantage of slicewise information. To this end, Çiçek et al. [11] proposed the 3-D U-Net to segment the target region from volume data. Since then, multiple 3-D segmentation approaches have been proposed based on the 3-D U-Net. Although these approaches achieved satisfactory results in some applications, they still have some obvious shortcomings, making them not clinically applicable in more challenging applications. First, due to the fixed kernel sizes and hence the limited receptive fields, most of these networks cannot effectively tackle scale variation, particularly when the variation is large. Second, some intrinsic characteristics in medical imaging, such as the various shapes of the tumor in different stages, the blurred boundaries caused by the invasion of adjacent organs, and the existence of noise and artifacts, make it difficult to achieve satisfactory segmentation results only relying on local information. However, it remains a challenging task to integrate effective multiscale schemes and nonlocal guidance mechanisms into a 3-D network without increasing computational complexity.

Many solutions have been proposed to overcome these shortcomings [12]. To tackle scale variation, Kamnitsas et al. [13] proposed parallel convolutional pathways to simultaneously incorporate local and global information, which significantly improves segmentation performance. Zhang et al. [14] applied multibranches with different dilated rate kernels and merge the output features to solve the limitation of insufficient receptive fields. Feng et al. [15] fused features generated from different stages using attention mechanisms to deal with scale variation. However, most of these methods treat features with different scales indiscriminately while it is more reasonable and effective to adjust the weights of different scales according to the inputs. Moreover, in the above schemes, it is usually hard to decide the best number of branches and dilated rates. Therefore, these schemes are still insufficient for segmentation tasks involving large-scale variation, such as kidney tumor segmentation.

On the other hand, to capture the long-range dependencies, many nonlocal guidance mechanisms have been proposed, aiming at leveraging nonlocal contexts to achieve better performance [16]. Li et al. [17] designed a self-attention module to model pixel relations regardless of their distance. Wang et al. [18] introduced a flexible global aggregation block into a 3-D U-Net to capture global dependencies by exploiting feature relationships. Recently, Xie et al. [19] applied a nonlocal mechanism in pulmonary lobe segmentation to assess pneumonia severity and progression for COVID-19 diagnosis and treatment. These works have demonstrated that nonlocal mechanism is able to dig more potentially useful features to further improve segmentation accuracy. However, while more or less improving the performance, these methods also inevitably introduced a lot of computation and memory costs, particularly in 3-D networks, which hampers them from being used in clinical settings without sufficient computing resources. To this end, how to effectively implement nonlocal context guidance (NCG) within a 3-D network remains a challenging issue.

In this article, we propose a 3-D CNN equipped with a new scale-aware feature extraction (SAFE) module and an efficient NCG for kidney and kidney tumor segmentation from Computed Tomography (CT) data. Different from existing solutions that deal with scale variation by directly fusing multiscale features, the proposed SAFE is capable of adaptively selecting appropriate receptive fields according to the inputs so that features of corresponding scales can be effectively filtered by assigning them with higher weights. Considering the segmentation accuracy of small tumors is one of the main challenges in this task, we further highlight the branch with small receptive fields when fusing the features. We further propose a novel yet efficient NCG to capture long-range dependencies to tackle irregular shapes and blurring boundaries of kidneys and tumors. Instead of directly harnessing a 3-D NCG, which makes the number of parameters exponentially increase and hence the network difficult to be trained under limited training data, we develop a 2.5-D NCG mechanism based on projections of feature cubes, which achieves a tradeoff between segmentation performance and network complexity. We extensively evaluate our *3DSN-Net* on the famous KiTS dataset with many challenging kidney and tumor cases. Experimental results demonstrate our solution consistently surpasses state-of-the-art 3-D networks equipped with scale aware and NCG mechanisms, particularly for tumor segmentation.

Our contribution can be summarized as follows.

1) We propose a novel 3-D network, called *3DSN-Net*, for kidney and kidney tumor segmentation from CT data, which is able to effectively tackle scale variation and ambiguous boundaries via adaptive scale-aware feature fusion and efficient NCG.
2) We propose an SAFE module to adaptively extract features according to the size of the target instead of indiscriminately fusing them regardless of the variation of the feature scale.
3) We present an efficient NCG mechanism without bringing a large number of parameters within a 3-D network as the previous 3-D nonlocal modules, maintaining a good balance between accuracy and efficiency.
4) We conducted extensive experiments to evaluate the proposed *3DSN-Net* on the famous public dataset KiTS. Experimental results show that our method achieves state-of-the-art performance and consistently outperforms other state-of-the-art 3-D networks, which demonstrates the advantages and effectiveness of our method.

The preliminary vision of this work has been published in [20]. In this article, we substantially revise the conference version to introduce the proposed *3DSN-Net* more thoroughly yet clearly. The main modifications include that: 1) we introduce the SAFE module to replace the depthwise separable convolutions in the conference version; 2) we have improved the methods with clearer elaborations for each step and added more detailed descriptions; and 3) we have conducted more ablation and comparison experiments using extra evaluation metrics, and comprehensively discussed both contributions and limitations of our *3DSN-Net*.

II. Related Work

A. Kidney and Kidney Tumor Segmentation

Most of the early studies on kidney and kidney tumor segmentation are developed based on hand-crafted features. For example, Gloger et al. [6] proposed a multistep refinement approach to produce probability maps and then applied an extended level-set model to improve the segmentation results. Xie et al. [5] first employed Gabor filters to extract texture features and then used an expectation-maximization method to construct texture models to segment kidneys from CT images. However, when the tumor shows a similar intensity with adjacent organs or has low contrast with the renal parenchyma, these hand-crafted features cannot produce satisfactory results to meet the clinical requirements. Recent works have demonstrated the superior performance of deep CNNs for medical imaging segmentation [9], [21], [22], [23], [24], [25], [26], including kidney and kidney tumor segmentation [27]. da Cruz et al. [28] first used AlexNet to produce a probability of whether a slice contains a kidney, then applied a modified 2-D U-Net for the kidney segmentation. Although this approach can improve the segmentation accuracy compared to traditional methods, it is difficult for a 2-D network to further improve the performance without considering the slicewise information. Recently, Yu et al. [29] have proposed a crossbar-Net, which consists of both vertical and horizontal feature encoder paths to simultaneously capture local and global knowledge of the kidney tumors from two perpendicular directions. This network achieves much more satisfactory results than its 2-D counterparts. However, how to efficiently extract and harness multiscale information and nonlocal information within a 3-D network remains a challenging task.

B. Multiscale Feature Extraction

As traditional 3-D U-Net is stacked by size-fixed convolutions, it is difficult to extract multiscale contextual information. Recently, to overcome this limitation, many works have been presented to fully exploit multiscale information [30], [31], [32]. One stream of the studies employed parallel pooling branches to capture multiscale features [33], [34], [35]. For example, PSPNet [36] and PoolNet [37] designed different pooling operations with various kernel sizes to generate features with rich regional contexts. The other stream of investigations considered using multibranch convolution with various kernel sizes to realize the multiscale feature extraction [38], [39]. For example, inception [40] and DeepLab [41] series designed different convolutional paths with various receptive fields to strengthen the capability to capture multiscale contextual knowledge. However, one main limitation of these existing solutions is that the kernel size or the number of branches needs to be fine-tuned to determine the best setting manually.

C. Nonlocal Mechanisms

Recently, nonlocal mechanisms have been widely applied to semantic segmentation tasks as they are able to capture long-range dependencies for better segmentation performance [42], [43], [44], [45], [46], [47]. Wang et al. [16] employed an attention mechanism to model relationships between each position and all other positions. Huang et al. [48] proposed an interlaced sparse nonlocal module to capture the global and local dependencies, respectively. Zhu et al. [49] designed a symmetric nonlocal block that integrates a pyramid sampling module into the nonlocal mechanism to boost its efficiency. Besides,the nonlocal mechanism has also been introduced in many realistic application scenarios [50], [51], [52], [53], [54]. For example, Xia et al. [50] proposed to encode nonlocal part-to-part correlations via second-order feature statistics for person reidentification. Mei et al. [54] presented a nonlocal sparse attention mechanism to retain long-range modeling capability using sparse representation for image super-resolution. However, most of these methods are designed and used in 2-D CNNs, and directly transferring them into 3-D CNNs may exponentially increase the number of the parameters, making

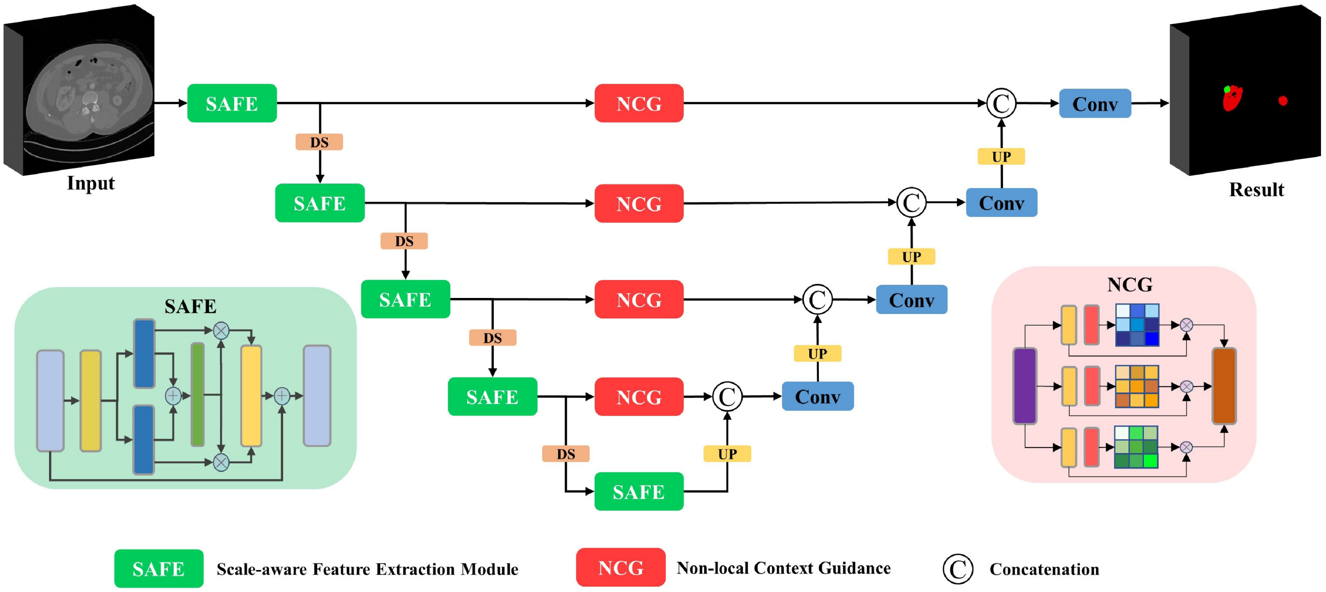

Input
SAFE
DS
NCG
Conv
UP
Result
SAFE
NCG
SAFE
Scale-aware Feature Extraction Module
NCG
Non-local Context Guidance
Concatenation

Fig. 2. Our proposed *3DSN-Net*. An SAFE module is equipped in the encoder to enrich receptive fields in different layers and enhance multiscale feature extraction, while an NCG module is employed in the skip connections to capture global context and fully exploit the long-range dependencies during the feature selections.

3-D networks even more difficult to be trained under limited training samples. In this regard, how to introduce nonlocal mechanisms into 3-D networks while effectively managing the computation and memory consumption requires careful studies.

III. Methodology

The architecture of our proposed *3DSN-Net* is illustrated in Fig. 2, which is developed based on a typical 3-D U-Net. The proposed *3DSN-Net* is composed of two key components. First, we propose a new 3-D SAFE module to adaptively select suitable receptive fields in different layers to deal with the targeting objects, particularly the tumors, with different sizes. Second, we develop a novel and efficient NCG module to exploit useful global dependencies based on the extracted scale-aware features to achieve segmentation performance improvement. Note that, we simply employ a 3-D transposed convolution operation to upsample features in decoder. Overall, we seamlessly combine the SAFE and NCG modules in the proposed *3DSN-Net* to systematically tackle the issues of the task.

A. Scale-Aware Feature Extraction

Traditional 3-D networks for medical image segmentation usually obtain a certain performance improvement by simply enlarging the size of convolution kernels or using a pyramid pooling scheme to roughly combine feature maps with different scales. These approaches, however, often require intensive experiments to determine the kernel size, the pyramid pooling size, and/or the weighting parameters within the networks, which is computation-intensive, time-consuming, and tedious, and hence prohibits them from being deployed in clinical practice.

We propose a new 3-D SAFE module to replace the traditional convolution block in 3-D U-Net (as shown in Fig. 3), which can capture multiscale information via multiple branches containing different kernel sizes. Based on the feature maps with multiscale features acquired by different kernel sizes, the proposed *3DSN-Net* dynamically determines the optimal kernel according to the size of the input kidney and tumor without using additional supervision or tedious parameter adjustment. Although the features extracted by branches in SAFE may be similar, the receptive field of the encoder is large enough with rich variations by stacking the SAFE modules across the contracting path of 3-D U-Net. When each SAFE module dynamically selects the optimal size of the convolution kernel, the entire encoder can naturally achieve the adaptive selection of receptive fields. It is worth noting that the SAFE can be conveniently applied in 2-D segmentation networks by simply replacing 3-D convolutions with 2-D convolutions.

For the given feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W\times S}$, where H, W, S and C denote height, width, slice, and channel, respectively, we first apply 3-D convolutions to downsample feature maps, and employ two $3 \times 3 \times 3$ kernels with different dilated rates ($r = 1, 2$) to capture multiscale features. When the size of the feature map is too small, applying convolution operation with a large dilated rate will probably cause the dilated convolution to lose its effect, so we choose a relatively small dilated rate here. By using the above dilated convolution operations, we can generate two feature maps, which can be defined as $\mathbf{U}_1$ and $\mathbf{U}_2$. We further fuse the two branches of features based on elementwise summation and obtain a feature map $\mathbf{U}$ containing multiscale information; it can be described as

$$\mathbf{U} = \mathbf{U}_1 + \mathbf{U}_2. \tag{1}$$

Then, we perform the kernel selection based on $\mathbf{U}$. To compact the global information for the kernel selection, we further conduct a global average pooling for each channel to obtain a global feature vector $\mathbf{p} \in \mathbb{R}^{C\times 1\times 1\times 1}$. Here, we can use $\mathbf{p}_c$ to

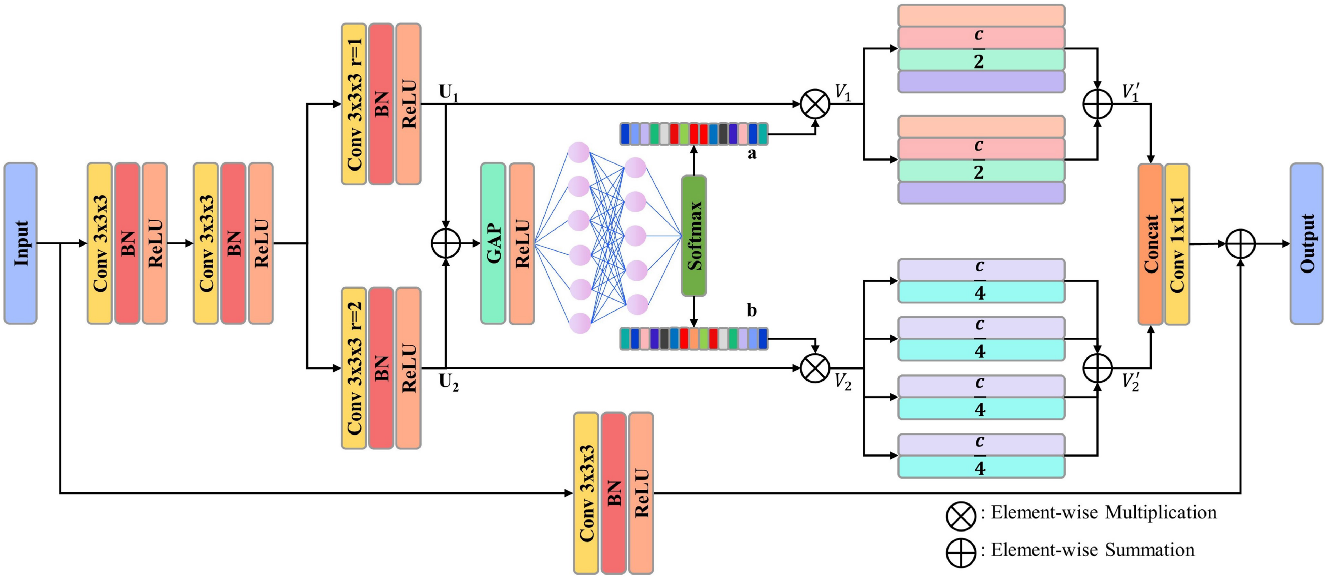


Fig. 3. SAFE module. Based on a multibranch architecture composed of various convolution kernel sizes, we can capture various receptive fields information and adaptively select the most suitable kernels according to the feedback of target sizes for the input kidney and tumors.

denote the cth element of $\mathbf{p}$ as

$$\mathbf{p}_c = F_{\text{gap}}(\mathbf{U}_c) = \frac{1}{H \times W \times S} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{S} \mathbf{U}_c(i, j, k) \tag{2}$$

where $\mathbf{U}_c$ is the cth channel of $\mathbf{U}$. Obviously, the size of $\mathbf{U}_c$ is $H \times W \times S$. Then we harness a simple fully connected layer to generate the guidance vector $\mathbf{z}$ for the kernel selection, which can be written as

$$\mathbf{z}_i = F_{fc}(\mathbf{p}) = W_i \delta(W_0(\mathbf{p})), \quad i \in \{1, 2\} \tag{3}$$

where W_0 and W_i are both linear transformation weights, and δ is the ReLU activation function. Considering the efficiency, we also use the reduction between two fully connected layers to minimize the parameters. The reduction rate can be defined as $d = \max(C/g, L)$, which controls the linear transformation matrices $W_0 \in \mathbb{R}^{d \times C}$, and $W_i \in \mathbb{R}^{C \times d}$. Increasing g can reduce the number of network parameters, but it also limits the representation capability of the fully connected layers. On the other hand, to prevent the generation of too small d when the number of feature channels is low, we set L to ensure that d is not less than this minimum value. In our experiments, we can set $g = 8$ and $L = 8$ to balance the accuracy and efficiency.

So far, we can obtain rough feature selection based on the guidance vector $\mathbf{z}_1$, $\mathbf{z}_2$ for two branches which are achieved by simple fully connected layers. To utilize the guidance vector $\mathbf{z}_1$, $\mathbf{z}_2$ for enabling the proposed *3DSN-Net* network to adaptively determine the optimal kernel sizes, we apply a softmax operator on the channelwise digits for highlighting the suitable receptive field from branches ($\mathbf{U}_1$ and $\mathbf{U}_2$), respectively. The normalized guidance vectors can be obtained as

$$[\mathbf{a}, \mathbf{b}] = \text{Softmax}(\text{Concat}(\mathbf{z}_1, \mathbf{z}_2)). \tag{4}$$

Hence, we obtain two weighted feature maps $\mathbf{V}_1 = \mathbf{a} \cdot \mathbf{U}_1$ and $\mathbf{V}_2 = \mathbf{b} \cdot \mathbf{U}_2$. Since the guidance vectors $\mathbf{a}$ and $\mathbf{b}$ are global and comprehensive representations generated from original feature maps $\mathbf{U}_1$ and $\mathbf{U}_2$ from two branches, we can adaptively achieve the selection of features with different receptive fields by multiplying it to $\mathbf{U}_1$ and $\mathbf{U}_2$. Thus, we can dynamically determine the optimal kernel size. As reported in GhostNet [55], in the training procedures, we also find that several extracted features in different channels could be very similar to each other. In this regard, to achieve a more compact representation, we further reduce the channel number by squeezing the feature maps $\mathbf{V}_1$ and $\mathbf{V}_2$. It is conducive to balancing the increase of parameters brought by multiple convolution branches, and simultaneously increasing the diversity and distinction of features across the channel. Specifically, we split the feature across the channel dimension into several features with the same number of channels, and compute elementwise summation of these features to obtain a low-channel feature. As feature maps extracted in the branch with smaller receptive fields contain more diverse details, we retain more channels for $\mathbf{V}_1$ in order to preserve these details. In our experiments, two more compacted feature maps $\mathbf{V}_1'$ and $\mathbf{V}_2'$ are generated by squeezing $\mathbf{V}_1$ and $\mathbf{V}_2$ to (1/2) and (1/4), respectively.

Finally, we figure out a feature map with rich multiscale information by concatenating the two branches $\mathbf{V}_1'$ and $\mathbf{V}_2'$. We further perform a 1×1 convolution to recover the feature number, making it equal to the original input. For a stable convergence, we further add a residual connection to generate the final output $\mathbf{V}$ of our SAFE module, which can be written as

$$\mathbf{V} = \text{Conv}_{1\times1}\big(\text{Concat}\big(\mathbf{V}_1', \mathbf{V}_2'\big)\big) + \mathbf{X}' \tag{5}$$

where $\mathbf{X}'$ denotes the residual part of the input feature $\mathbf{X}$.

B. NCG Mechanism

Based on a 3-D network equipped with SAFE, we further propose a new yet efficient NCG mechanism. Due to the limited sizes of convolutional kernels, the features can only capture local information between adjacent voxels, which is incapable to extract valuable long-range information, making it insufficient to tackle challenging cases with irregular shapes and blurred boundaries. To this end, we develop a nonlocal attention mechanism to capture global long-range dependencies to enrich the feature representation ability. Based on the proposed nonlocal attention mechanism, we can extract the key long-range yet high-level dependencies to achieve segmentation performance improvement. The main challenge of integrating nonlocal information in our 3-D network is that considering the limited computational resources in clinical settings, we cannot directly extend existing 2-D nonlocal modeling techniques to 3-D versions, and apply them to the 3-D kidney and tumor segmentation, as the parameters will exponentially increase, which makes the 3-D network difficult, if not possible, to be effectively trained. In this regard, we introduce a somehow 2.5-D nonlocal attention mechanism based on feature projections from three directions. Actually, our NCG module generates the attention weights by multiplying the query features with the normalized key features, and then employs the attention weights to achieve different degrees of activation within features from three projection directions. Therefore, our NCG is essentially an attention module. Note that our NCG module is specially designed for 3-D networks, which is not suitable for 2-D segmentation tasks.

The proposed NCG module is illustrated in Fig. 4. Concretely, we feed our NCG with a feature map $\mathbf{U} = \{\mathbf{U}_1, \ldots, \mathbf{U}_C\}$, whose size is $H \times W \times S$ with $H = W$ for a cube volume. We then project the feature cube (3-D feature maps) into three projection views, which can be formulated as

$$\mathbf{V}_h(i) = \frac{1}{H}\sum_{i=1}^{H}\mathbf{U}_c(i,j,k) \tag{6}$$

$$\mathbf{V}_w(j) = \frac{1}{W}\sum_{j=1}^{W}\mathbf{U}_c(i,j,k) \tag{7}$$

$$\mathbf{V}_s(k) = \frac{1}{S}\sum_{k=1}^{S}\mathbf{U}_c(i,j,k). \tag{8}$$

By projecting 3-D feature maps into 2-D, three feature maps are generated, including $\mathbf{V}_h \in \mathbb{R}^{C\times W\times S}$, $\mathbf{V}_w \in \mathbb{R}^{C\times H\times S}$, and $\mathbf{V}_s \in \mathbb{R}^{C\times H\times W}$. To encode the pixelwise correlation, we compute the pixel relationship vector $\mathbf{q} \in \mathbb{R}^{C\times 1\times 1}$ within the 2-D feature maps as

$$\mathbf{q}_{t_c} = \sum_{j=1}^{N_t}\frac{e^{(W\mathbf{V}_{t_c}(j))}}{\sum_{m=1}^{N_t} e^{(W\mathbf{V}_{t_c}(m))}}\mathbf{V}_{t_c}(j), \quad t \in \{h, w, s\} \tag{9}$$

where t denotes three different dimensions of h, w, and s, respectively, and $N_h = S \times W, N_w = S \times H, N_s = H \times W$, respectively. After performing three equal transforming operations across three different dimensions, we employ a fully connected layer to obtain more nonlinear features, which can be written as

$$\mathbf{z}_t = \tilde{\mathbf{F}}_{fc2}(\mathbf{q}_t) = W_4\delta(W_3(\mathbf{q}_t)) \tag{10}$$

where δ is the ReLU activation function and $\mathbf{z}_t \in \mathbb{R}^{C\times 1\times 1}$. Likewise, we also reduce the model parameters via a reduction ratio g set to 8. Given the $\mathbf{z}_t \in \mathbb{R}^{C\times 1\times 1}$. We then project $\mathbf{z}_t$ to three different dimensions and fuse the features via pixelwise summations. Additionally, we further prevent network degradation via the addition of a residual connection

$$\mathbf{O}_t = \mathbf{z}_t \cdot \mathbf{V}_t + \mathbf{V}_t, \quad t \in \{h, w, s\}. \tag{11}$$

Ultimately, we employ dimension broadcasting elementwise summation operation (i.e., sum fusion shown in Fig. 4) to fuse three 2-D feature maps $\mathbf{O}_h$, $\mathbf{O}_w$, and $\mathbf{O}_s$ into a 3-D feature map, whose size is the same as the original volume.

C. Loss Function

In the task of kidney and kidney tumor segmentation, the small target is one of the main challenges. Although Dice loss is widely used for medical image segmentation, it is unstable when dealing with small target segmentation. When the small target has several pixel misclassifications, it will cause significant changes in Dice, resulting in dramatic gradient changes and unstable training. To achieve a stable convergence during training, we employ a combination of cross-entropy [56] and Dice loss [57] with the same weighting

$$L = -\frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N}\left(y_{n,c}\log p_{n,c} + \frac{2y_{n,c}p_{n,c}}{y_{n,c}^2 + p_{n,c}^2}\right) \tag{12}$$

where $y_{n,c}$ and $p_{n,c}$ denote the target label and prediction probabilities for class c of the nth pixel. Moreover, similar to UNet++ [58], we utilize deep supervision to make the network converge more stably. Concretely, we apply multiple segmentation heads on features of each stage of our networks to obtain outputs of the corresponding stage. Then we compute the losses between outputs of each stage and ground truths to supervise networks with different semantic scales. The overall loss is the weighted summation of all losses, whose weights are set as in UNet++ [58].

IV. Results

A. Dataset and Implementation Details

We conducted extensive experiments to verify the proposed *3DSN-Net* on the famous KiTS dataset [59], which includes 210 3-D CT data scans. We randomly selected 42 scans in the KiTS dataset as validation set, and train the model on the remaining 168 CT scans. As the images have different sizes and spacings, we resampled the images to an isotropic and uniform spatial resolution of 192 × 192. At the same time, we clipped the CT intensity values to fall into the range of [1%, 99%] percentile and normalized them to [–1, 1]. We performed three different kinds of data augmentations, including random flipping, randomly rotating, and adding noise. To do so, we can obtain a more diverse dataset for a more comprehensive evaluation of kidney and tumor segmentation. Moreover, to evaluate the generalization of our method, we also applied our method on the LiTS dataset [60], consisting of 131 annotated 3-D CT data scans. We randomly selected 26 scans in the LiTS dataset as validation set, and train the model on the remaining 105 CT scans. In our experiments, we also performed resampling and data augmentations similar to the KiTS dataset.

We initialize our networks via the Kaiming initialization method [61]. Additionally, we apply Adam optimizer [62] during the training process, where the initial learning rates are

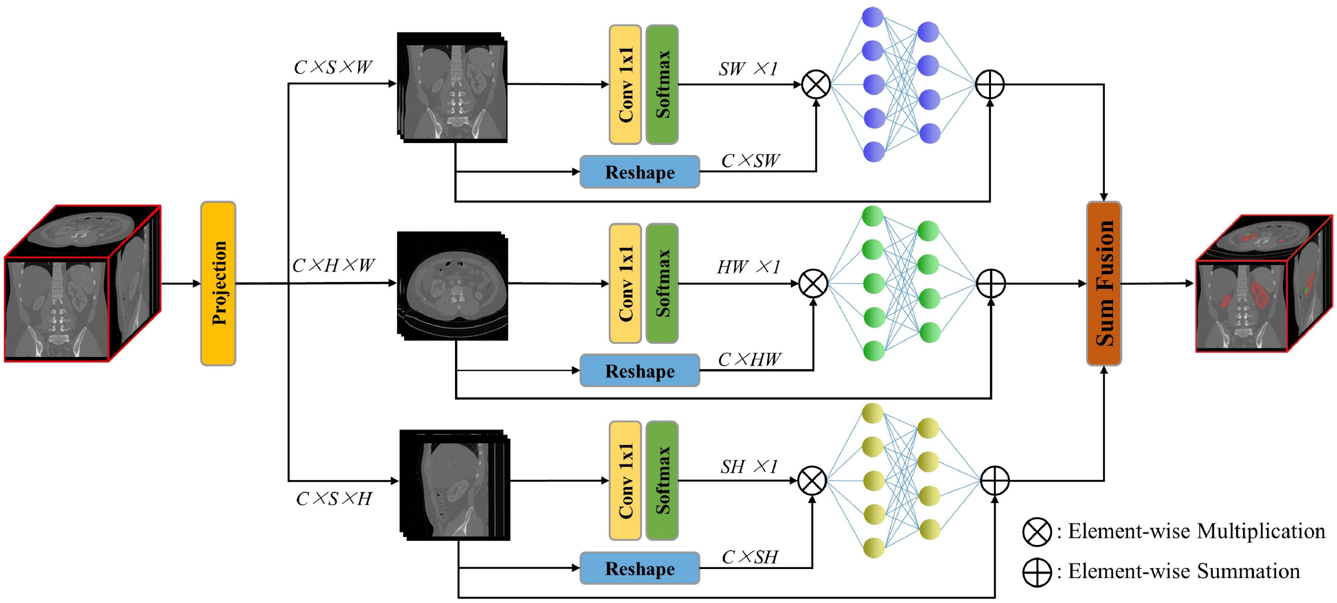


Fig. 4. Proposed NCG module. To effectively capture long-range dependencies among high-level features while suppressing the worthless local redundant relationships, a 2.5D nonlocal mechanism based on projections from three directions is developed.

set to 0.001 for the KiTS dataset and 0.0005 for the LiTS dataset with the Cosine Annealing scheduling strategy. In our experiments, we set the batch size to 2 and train our proposed *3DSN-Net* for 600 epochs.

B. Evaluation Metrics

We employed four metrics to verify the effectiveness of the proposed method, including volume-based metrics and surface distance-based metrics. The volume-based metrics include Dice coefficient (DICE), and relative absolute volume difference (RAVD) which can be written as

$$\mathrm{DICE} = \frac{2|P \cap Y|}{|P| + |Y|} \tag{13}$$

$$\mathrm{RAVD} = |\frac{|P| - |Y|}{|Y|}| \tag{14}$$

where P and Y denote the prediction of the network and ground truth, respectively. DICE is calculated based on the overlapping volume between prediction and ground truth, while RAVD is calculated based on the overlapping area between prediction and ground truth.

On the other hand, to meet the requirement of real applications, we also use surface distance-based measures metrics, such as average symmetric surface distance (ASSD), and 95% Hausdorff Distance (95HD, in voxel), which are crucial for surgical operations. ASSD and 95HD can be formulated as follows:

$$\mathrm{ASSD} = \frac{\sum_{x \in B_P} \mathrm{dis}(x, B_G) + \sum_{y \in B_G} \mathrm{dis}(y, B_P)}{|B_P| + |B_G|} \tag{15}$$

$$\mathrm{dis}(x, A) = \min_{y \in A} \mathrm{dis}(x, y) \tag{16}$$

where B_P and B_G denote the boundary of P and G, respectively. $\mathrm{dis}(x, A)$ represents the distance from voxel x to voxel set A, where $\mathrm{dis}(x, y)$ denotes the Euclidean distance. The HD can measure the symmetric surface distance between the ground truth and prediction, which can be written as

$$\mathrm{dis}_H(A, B) = \max_{x \in A} \min_{y \in B} \mathrm{dis}(x, y) = \max_{x \in A} \mathrm{dis}(x, B) \tag{17}$$

$$\mathrm{HD} = \max\{\mathrm{dis}_H(B_P, B_G), \mathrm{dis}_H(B_G, B_P)\} \tag{18}$$

where $\mathrm{dis}_H(A, B)$ is used to define the maximal distance between two sets A and B by calculating a point in the first to the nearest point in the other one. Note that, we used the 95th percentile of the asymmetric HD instead of the maximum to alleviate the effects of outliers.

C. Ablation Studies

1) Ablation Studies on SAFE and NCG: To verify the effectiveness of the SAFE and NCG modules on the proposed *3DSN-Net*, we performed a series of ablation studies on the KiTS dataset. The 3-D U-Net is used as the baseline in our ablation studies. By adding different components to the baseline, we can quantitatively verify the effectiveness of the SAFE and NCG.

Some typical cases of our ablation studies are shown in Fig. 5. Fig. 5(d) illustrates that the 3-D U-Net failed to handle the challenging cases with various scales, irregular shapes, and blurring boundaries. By adding the NCG module, which aims to capture long-range dependencies to achieve the segmentation performance improvement, Baseline+NCG method outperforms the 3-D U-Net in handling kidney and tumors with blurring boundaries [as shown in Fig. 5(e)]. By adopting suitable scales for kidneys and tumors and encoding more effective features using the SAFE module, Baseline+SAFE method obtains better segmentation accuracy than the baseline 3-D U-Net, particularly for the cases with various scales (as shown in column (f) of Fig. 5). By seamlessly integrating SAFE and NCG in the proposed *3DSN-Net*, we can not only extract richer multiscale features but also capture

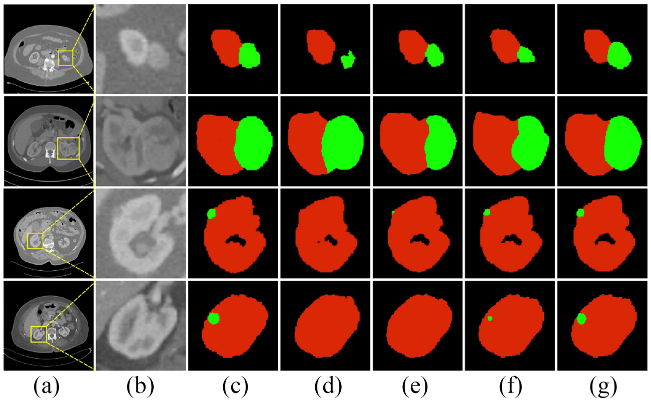


Fig. 5. Visual comparisons of ablation studies. Red and green regions indicate kidney and tumors, respectively. (a) Input. (b) Blow up. (c) Ground truth. (d) Baseline (3-D U-Net). (e) Baseline + NCG. (f) Baseline + SAFE. (g) Baseline + NCG + SAFE.

TABLE I
STATISTICAL COMPARISONS OF ABLATION STUDY ON SAFE AND NCG MODULES (IN MEAN±SD)

Method	DC (kidney)	DC (tumor)	RAVD	ASSD	95HD
Baseline	0.963±0.02	0.806±0.39	0.11±0.48	0.61±0.81	3.25±2.57
Baseline+SAFE	0.969±0.02	0.852±0.23	0.04±0.12	0.43±0.27	1.83±1.36
Baseline+NCG	0.971±0.01	0.849±0.31	0.05±0.15	0.39±0.11	2.23±1.71
Baseline+SAFE+NCG	**0.975±0.01**	**0.872±0.17**	**0.02±0.10**	**0.33±0.14**	**1.40±1.18**

more long-range dependencies, which makes our *3DSN-Net* outperform other competitors in these challenging cases, as shown in Fig. 5(g).

Furthermore, we conducted a quantitative comparison by gathering the mean Dice (kidney), Dice (tumor), RAVD, ASSD, and 95HD values of different approaches tested on the KiTS validation dataset. As illustrated in Table I, where the best results are highlighted in bold, we can easily observe that both Baseline+SAFE and Baseline+NCG methods are superior to the conventional 3-D U-Net, which clearly demonstrates the effectiveness of our SAFE and NCG modules. We can also observe the complementariness of the SAFE and the NCG modules. On the one hand, in terms of volume-based metrics (Dice and RAVD), the Baseline+SAFE obtains better performance than Baseline+NCG, indicating the capability of the SAFE module in tackling the scale variation for covering more areas of the targets. On the other hand, in terms of distance-based metrics (ASSD and 95HD), the Baseline+NCG generally outperforms the Baseline+SAFE, indicating the capability of the NCG module in dealing with blurring boundaries by capturing more long-range dependencies. As shown in the bottom row of Table I, the proposed *3DSN-Net* obtains the best statistical performance for all 5 evaluation metrics in the ablation studies, demonstrating that the integration of SAFE and NCG modules is able to reinforce each other and achieves better segmentation results.

2) Ablation Studies Inside SAFE: To verify the effectiveness of different components within the SAFE module, we further conducted ablation experiments inside the SAFE. As the proposed SAFE module can be easily expanded into more branches, we first conducted an experiment to evaluate the impact of branch numbers on the segmentation accuracy of the SAFE module. Concretely, we compared the performance of 3-branch SAFE and 2-branch SAFE. In the experiment, 3-branch SAFE is also based on the same convolution kernel size but with different dilated rates (r = 1, 2, 4) in different branches, respectively. As shown in Table II, we can observe that experimental results that the accuracy of 3-branch SAFE is lower than that of 2-branch SAFE, which indicates that two branches are sufficient in dealing with the task of kidney and tumor segmentation. In theory, the performance of the 3-branch SAFE should be better than that of the 2-branch SAFE, because the number of parameters of the model increases, and the additional dilated rate branch increases the selection space of the receptive fields. However, the amount of samples in medical image datasets is usually limited, and the dataset KiTS used in the experiment only contains 210 cases of 3-D CT scans, which easily leads to overfitting of the model. That is why we obtain a relatively poor performance for the 3-branch SAFE. Our experimental results in Table III show that, whether using (1, 2) scheme or (1, 6) scheme, adding an extra branch not only fails to improve the segmentation performance but causes a slight decline in accuracy, which indicates that adding redundant branches increases the complexity of the model and exacerbates the risk of model overfitting, resulting in a decrease in segmentation accuracy. We further compared the segmentation performances between the 2-branch SAFE and 3-branch SAFE on the additional LiTS dataset [60]. Experimental results show that the segmentation Dice for liver and tumor drops from 0.968 and 0.668 to 0.965 and 0.661, further demonstrating that the 3-branch scheme is likely to bring model overfitting for our 3-D network architecture. Based on the experimental results, we used 2-branch SAFE in the proposed *3DSN-Net*. We also conducted an ablation study based on the 2-branch SAFE to validate the effectiveness of the channel squeezing. As shown in Table II, we can clearly see that SAFE with channel squeezing generally outperforms the SAFE without it in all five metrics. As pointed out by GhostNet [55], there exists a number of redundant channels with very similar extracted features in traditional networks. By squeezing the channels in our extracted feature maps, we can obtain a more compact feature representation. More importantly, as the channel squeezing can also enhance the distinction of each channel in a diverse feature map, we can hence improve segmentation performance in kidney and tumor segmentation. Besides, we

TABLE II
STATISTICAL COMPARISONS OF ABLATION STUDY INSIDE SAFE MODULE (IN MEAN±SD)

Method	DC (kidney)	DC (tumor)	RAVD	ASSD	95HD
Baseline	0.963±0.02	0.806±0.39	0.11±0.48	0.61±0.81	3.25±2.57
Baseline+2-branch SAFE	**0.975±0.01**	**0.872±0.17**	**0.02±0.10**	**0.33±0.14**	**1.40±1.18**
Baseline+3-branch SAFE	0.973±0.01	0.868±0.22	0.02±0.14	0.33±0.16	1.46±1.24
Baseline+SAFE w/o sqz	0.968±0.01	0.863±0.20	0.03±0.15	0.36±0.18	1.56±1.44
Baseline+AtrousConv	0.963±0.01	0.841±0.23	0.04±0.23	0.51±0.36	2.03±1.66

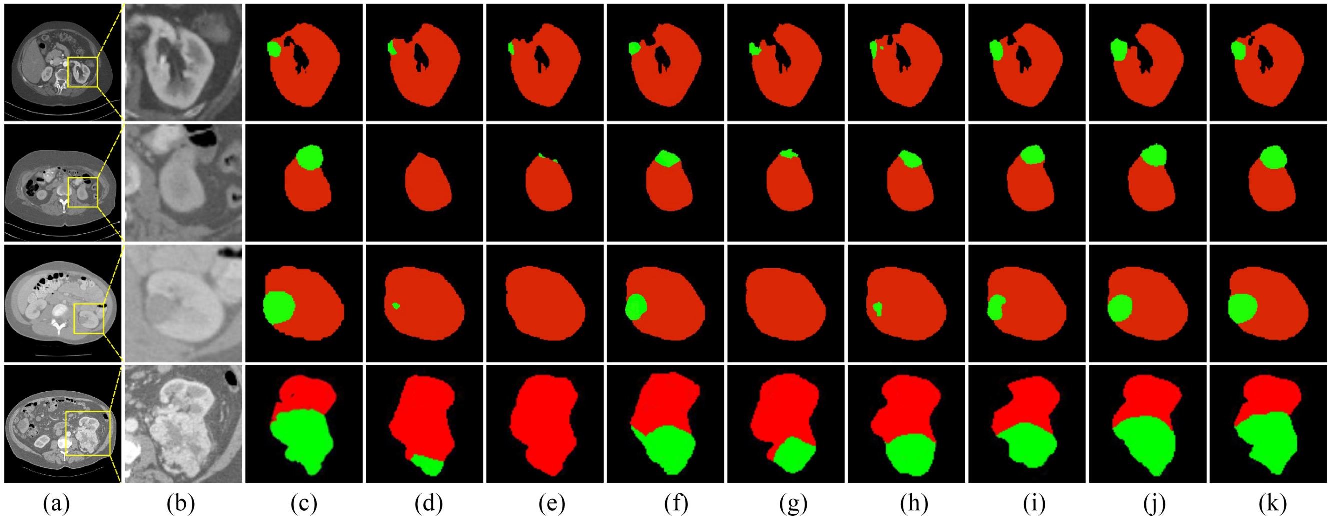


Fig. 6. Visual comparisons with different state-of-the-art methods. Red and green regions represent kidneys and tumors, respectively. The size of tumors gradually increases from top to bottom. (a) Input. (b) Blow up. (c) Ground truth. (d) 2.5D P-UNet. (e) 3-D U-Net. (f) 3-D nnU-Net. (g) 3-D Res-UNet. (h) 3-D Attention-UNet. (i) 3-D Dense-VNet. (j) 3-D Nonlocal UNet. (k) Proposed *3DSN-Net*.

TABLE III
STATISTICAL COMPARISONS OF ABLATION STUDY ON DILATED RATES (IN MEAN±SD)

Dilated rates	DC (kidney)	DC (tumor)	RAVD	ASSD	95HD
(1,2)	**0.975±0.01**	**0.872±0.17**	**0.02±0.10**	**0.33±0.14**	**1.40±1.18**
(1,2,3)	0.971±0.01	0.870±0.19	0.02±0.12	0.34±0.17	1.43±1.32
(1,2,4)	0.973±0.01	0.868±0.22	0.02±0.14	0.33±0.16	1.46±1.24
(1,6)	0.968±0.02	0.867±0.18	0.03±0.15	0.34±0.15	1.45±1.39
(1,6,12)	0.966±0.02	0.861±0.25	0.04±0.19	0.38±0.19	1.55±1.48

also replaced the SAFE modules with atrous convolutions with dilated rate 2 to form a new competitor Baseline+AtrousConv, as shown in Table II. We can obviously obverse that the segmentation performance of this method is significantly worse than those using SAFE, regardless of whether SAFE uses 2-branch, 3-branch or has no channel squeezing. The experimental results further demonstrate the effectiveness of our proposed SAFE module.

3) Ablation Studies on Dilated Rates: To further analyze the effect of different combinations of dilated rates in different branches of the SAFE module, we also performed ablation experiments with different dilated rates. Concretely, we experimentally adjust the dilated rates and observe their performance changes. In the experiment, we test the performance of schemes with dilated ratios of $(1, 2)$, $(1, 2, 3)$, $(1, 2, 4)$, $(1, 6)$, and $(1, 6, 12)$, respectively, where the $(1, 2)$ scheme is our adopted strategy. As shown in Table III, we can observe that we can obtain relatively better results when the dilated rates of $(1, 2)$ are applied in SAFE. After we add a branch, we can clearly see a drop in segmentation performance in most metrics, regardless of whether the dilated rate of the additional branch is 3 or 4, although the decline is not significant. As mentioned before, adding redundant branches not only fails to improve the segmentation performance but increases the complexity of the model and exacerbates the risk of model overfitting. Moreover, with the dilated rates of $(1, 6)$, we obtain worse performance than the $(1, 2)$ scheme, although the $(1, 6)$ scheme theoretically achieves a larger receptive field. We believe that applying convolution operation with a large dilated rate probably causes the dilated convolution to lose its effect when the size of the feature map is too small, because the range of dilated convolution at most positions crosses the boundary of the feature map. The segmentation performance with dilated rates $(1, 6, 12)$ naturally declines further, because the features of the extra branches are likely to have little feature scale variety due to the excessive dilated rate of 12.

D. Comparisons With State-of-the-Art Methods

We compared the proposed *3DSN-Net* with six state-of-the-art 3-D segmentation networks, including 3-D U-Net [11], 3-D Res-UNet [63], 3-D Att-UNet [64], Dense V-Net [65], Nonlocal UNet [18], and one-stage 3-D nnU-Net [66], and a 2.5D segmentation network P-UNet [67], to further evaluate the performance of the proposed approach in kidney and tumor segmentation. We implemented the seven competitors based on the same preprocessed dataset produced by nnU-Net framework [66]. For a fair comparison, both qualitative and quantitative results are collected using identical data augmentations, and under identical dataset settings and computational environments.

Fig. 6 shows the segmentation results of some typical challenging cases yielded by the seven state-of-the-art methods as well as our approach. By projecting 3-D images from three directions, the 2.5-D P-UNet fuses the segmentation results of three perpendicular 2.5-D Res-UNets. However, limited to the capability of the 2.5-D segmentation network, it is still difficult to obtain satisfactory segmentation results. Although 3-D networks can extract information from 3-D images more efficiently, there also have their own limitations. It is clearly observed that the 3-D U-Net cannot handle kidneys and tumors appearing with relatively small scales, irregular

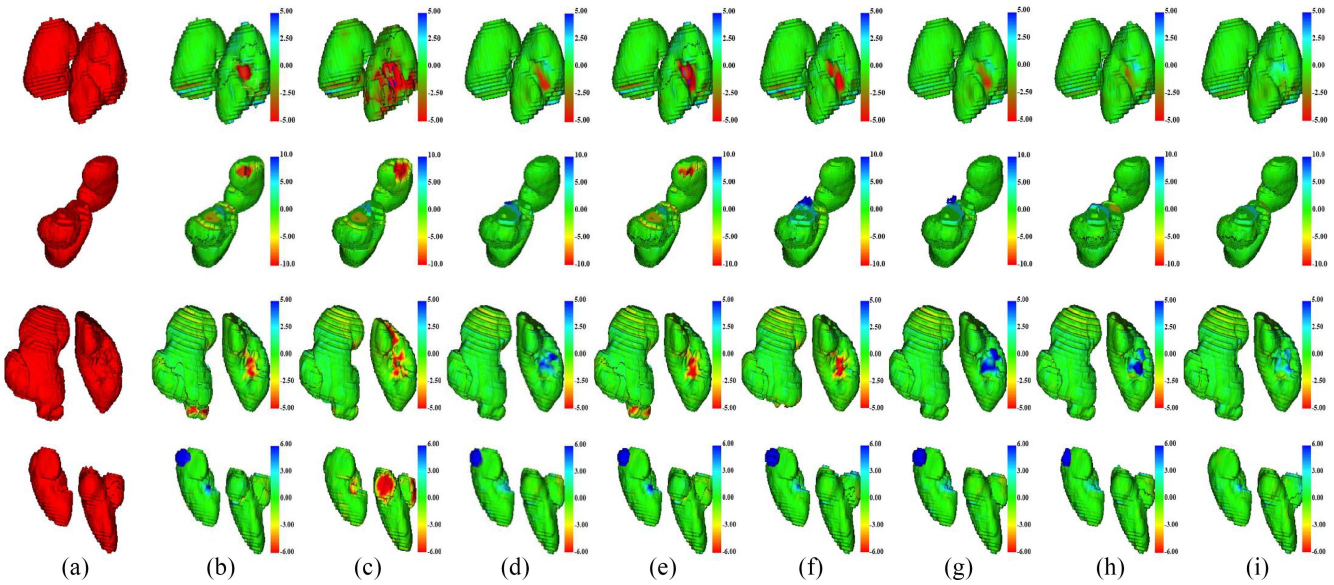


Fig. 7. Visual surface distance comparisons with different state-of-the-art methods: (a) ground truth, (b) 2.5D P-UNet, (c) 3-D U-Net, (d) 3-D nnU-Net, (e) 3-D Res-UNet, (f) 3-D Attention-UNet, (g) 3-D Dense-VNet, (h) 3-D Nonlocal UNet, and (i) proposed *3DSN-Net*.

shapes, or blurring boundaries. One-stage 3-D nnU-Net suffers from similar problems, although it achieved better segmentation performance than 3-D U-Net benefited from its network configurations. Thanks to a set of residual structure blocks, which can extract more representation features than traditional convolutional blocks, the 3-D Res-UNet gained better segmentation performance than the 3-D U-Net. However, it still cannot tackle these challenging cases due to the limited receptive fields. By adding a gated attention mechanism, the 3-D Attention-UNet can effectively select distinguishing features, and hence outperform the 3-D U-Net and the 3-D Res-UNet. However, the gated attention mechanism still cannot capture long-range global information. By introducing a set of dense blocks, 3-D Dense-VNet further enhanced the kidney and tumor segmentation with a much more powerful encoder. Similar to the previous methods, it also suffered from lacking long-range global information, resulting in unsatisfactory results in challenging cases with large/small scales and irregular shapes. More recently, based on a nonlocal mechanism for weighting the features with global contexts, 3-D nonlocal U-Net further improved the tumor segmentation accuracy with a global aggregation block. However, without effectively aggregating multiscale receptive fields in the feature extractions and efficiently extracting nonlocal long-range global contexts, all the above competitors still cannot obtain a satisfied kidney and tumor segmentation performance in many challenging cases. By seamlessly integrating the SAFE and NCG in the proposed *3DSN-Net*, we can observe that our method generally surpasses the seven competitors in the challenging cases with varying scales, irregular shapes, and blurring boundaries for the kidney and tumors (as shown in the last column of Fig. 6), manifesting the effectiveness of combining the SAFE and NCG for kidney and tumor segmentation. The Fig. 6 shows that our method can deal with tumors of varying scales, while other competitors tend to predict the size of the tumor to be smaller when the tumor boundaries are blurred. Additionally, we also visualize the corresponding surface distance between the prediction of each comparison method and the ground truth, as shown in Fig. 7. It is also observed that our approach consistently obtains higher-segmentation performance than the other seven competitors.

TABLE IV
STATISTICAL COMPARISONS WITH DIFFERENT STATE-OF-THE-ART METHODS (IN MEAN±SD)

Method	DC (kidney)	DC (tumor)	RAVD	ASSD	95HD
2.5D P-UNet	0.962±0.01	0.817±0.18	0.13±0.34	0.65±0.62	3.10±1.94
3D U-Net	0.963±0.02	0.806±0.39	0.11±0.48	0.61±0.81	3.25±2.57
3D nnU-Net	0.968±0.01	0.852±0.25	0.03±0.19	0.45±0.18	2.03±1.80
3D Res-UNet	0.968±0.01	0.831±0.33	0.08±0.20	0.51±0.38	2.32±2.08
3D Att-UNet	0.969±0.02	0.834±0.27	0.06±0.16	0.46±0.56	2.20±2.07
3D Dense-VNet	0.967±0.01	0.853±0.21	0.04±0.13	0.48±0.32	1.87±1.91
3D NL UNet	0.973±0.01	0.869±0.23	0.02±0.15	0.38±0.20	1.69±1.53
Ours	**0.975±0.01**	**0.872±0.17**	**0.02±0.10**	**0.33±0.14**	**1.40±1.18**

We further performed statistics on the average Dice (kidney), Dice (tumor), RAVD, ASSD, and 95HD values of different approaches evaluated on the KiTS validation dataset. Notice that all the statistical comparisons are based on five-fold cross-validation. Table IV shows the statistical results for all competitors. For both volume-based metrics or surface distance-based metrics, it is obviously observed that our method consistently surpasses the other seven competitors, achieving the best kidney Dice value of 0.97, tumor Dice of 0.87, RAVD of 0.02 voxels, ASSD of 0.33 voxels, and 95HD of 1.40 voxels. Compared with the 3-D U-Net that is widely used for medical imaging segmentation, our *3DSN-Net* obtained 1.1%, 6.5%, 9%, 45.9%, and 56.9% improvements in the metrics of kidney DICE, tumor DICE, RAVD, MSSD, and 95HD, respectively, indicating the effectiveness of our

TABLE V
STATISTICAL COMPARISONS WITH DIFFERENT STATE-OF-THE-ART METHODS ON LITS DATASET (IN MEAN±SD)

Method	DC (liver)	DC (tumor)	RAVD	ASSD	95HD
2.5D P-UNet	0.950±0.02	0.612±0.38	0.32±0.25	3.57±0.94	11.89±2.06
3D U-Net	0.949±0.02	0.591±0.29	0.31±0.33	3.61±0.63	12.66±2.99
3D nnU-Net	0.957±0.02	0.637±0.22	0.23±0.27	2.51±0.59	7.91±2.15
3D Res-UNet	0.954±0.02	0.630±0.47	0.31±0.20	3.02±1.31	9.03±3.42
3D Att-UNet	0.953±0.02	0.624±0.37	0.25±0.21	2.73±0.87	10.65±2.79
3D Dense-VNet	0.956±0.01	0.657±0.20	0.24±0.35	2.84±0.62	7.28±2.54
3D NL UNet	0.960±0.01	0.649±0.36	0.19±0.29	2.06±0.85	6.58±2.00
Ours	**0.968±0.01**	**0.668±0.21**	**0.15±0.18**	**1.58±0.39**	**6.37±2.02**

TABLE VI
COMPLEXITY COMPARISONS WITH STATE-OF-THE-ART METHODS

Method	Params (M)	FLOPs (G)	Mem (MB)	Time (ms)
2.5D P-UNet	49.5	330	969	81
3D U-Net	21.3	158	**370**	52
3D nnU-Net	24.1	162	464	54
3D Res-UNet	30.0	203	648	77
3D Att-UNet	25.9	173	475	57
3D Dense-VNet	26.3	256	882	106
3D NL UNet	**18.8**	165	615	62
Ours	19.1	**138**	403	**51**

SAFE and NCG modules in capturing multiscale features and harnessing the global contexts.

To validate the generalization capability of our proposed method, we have further applied our method on the LiTS dataset for liver and tumor segmentation and conducted the statistical comparison experiment to compare our method with other state-of-the-art methods based on five-fold cross-validation. As shown in Table V, we can obtain 0.96 in liver Dice, 0.66 in tumor Dice, 0.15 in RAVD, 1.58 voxels in ASSD and 6.37 voxels in 95HD, which outperforms other competitors in all 3-D segmentation metrics. The experimental results further demonstrate the generalization of our network.

On the other hand, we also compare our approach with other competitors in terms of parameter amounts, floating-point operations per second (FLOPs), memory overhead, and inference time, to analyze the complexity. The collected parameter amounts, FLOPs, memory overhead, and inference time for different competitors are shown in Table VI. Different from other approaches that usually exchange the performance gains with computational resources expense, our method only requires 19.1M parameters, 138 GFLOPs, 403-MB memory, and 51-ms inference time. Obviously, our approach gains better statistical performance than the other seven competitors in terms of FLOPs and inference time. Compared with other state-of-the-art methods, our approach not only consistently surpasses the competitors with regard to segmentation accuracy but also preserves a relatively efficient network, which is conducive to model deployment considering the limited computational resources in clinical practice.

TABLE VII
STATISTICAL COMPARISONS WITH STATE-OF-THE-ART METHODS IN THE CHALLENGE LEADERBOARD

Method	Mean DC	DC (kidney)	DC (tumor)
Isensee	0.9168	0.9793	0.8542
junma	0.9147	0.9738	0.8555
GrandBeam	0.9138	0.9787	0.8489
SZU (Ours)	*0.9127*	*0.9739*	*0.8515*
bubblyyi	0.9112	0.9774	0.8451
HeliXes	0.9089	0.9728	0.8450
PingAnTech	0.9064	0.9674	0.8454

E. Comparisons With Multistage Methods in the Challenge Leaderboard

We have submitted our results obtained from the proposed approach of this revised paper, and compared our approach with the multistage methods reported in the kidney and kidney tumor segmentation competition leaderboard (as shown in Table VII). For the KiTS2019 competition leaderboard,[1] we found that all solutions are ranked based on mean kidney tumor dice. We can clearly observe that only little differences appear in the dice of kidneys among different methods. However, for the segmentation accuracy of kidney tumors, the gap among different methods has gradually widened, which is more valuable for clinical applications. Although our method is a single-stage method, our method still ranked third in tumor dice of 0.8515, which is only 0.4% behind the first place. Obviously, multistage schemes might be more competitive in tumor region refinement, but they usually also require much heavier parameters, which also demand much more computational resources. Based on only a single-stage scheme, our method can efficiently integrate the SAFE module and the NCG module into the 3-D U-Net to achieve accurate kidney tumor segmentation. Compared with the multistage schemes, which commonly require more than one model for inference, our method is much more lightweight and memory efficient. More importantly, our method is an end-to-end method, which is also more suitable for clinical applications.

F. Discussion and Limitation

From the comparative experiments with state-of-the-art approaches, existing approaches cannot achieve satisfactory segmentation results in this task due to two major challenges of various scales and blurred tumor boundaries in kidney and tumor segmentation. Although recent approaches propose many PPM-like or multibranch schemes to capture multiscale features to deal with various scales problem, these methods need carefully designed to fit different situations. Therefore, we design an SAFE module that can adaptively select the optimal receptive field from multiscale context information to segment various scales of kidneys and tumors. Moreover, to better handle the situations of boundaries blurred and adjacent problems, we introduce the nonlocal mechanism to extract global long-range dependencies to highlight the target region

[1] https://kits19.grand-challenge.org/evaluation/challenge/leaderboard

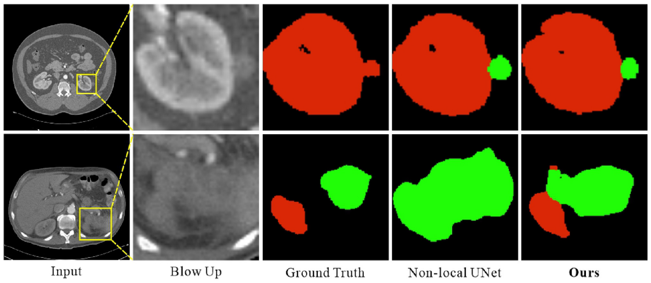


Fig. 8. Failure cases. Red and green regions indicate kidney and tumors, respectively.

with a global perspective. By seamlessly integrating SAFE and NCG modules, our SNCG-Net can not only fully capture multiscale features but also extract richer nonlocal context, which is proved to be very effective for the challenging kidney and tumor segmentation task.

Although our approach outperforms the state-of-the-art approaches, the proposed *3DSN-Net* still failed to deal with some extremely challenging cases. For the cases where even humans still cannot distinguish the kidney cysts from kidney tumors, or the cases, including very large tumors accompanied by other organ adhesions (as shown in Fig. 8), our *3DSN-Net* still cannot accurately segment the kidney and tumors. As our training data only contains very limited cysts cases, it is almost impossible for the network to distinguish the characteristics of cysts from real tumors. For the cases with different organ adhesions, we can further expand the receptive fields and make full use of nonlocal information to improve the segmentation performance. From our studies, we find that effectively and efficiently harnessing long-range dependencies in 3-D networks is a potential way to improve segmentation performance. As the proposed *3DSN-Net* can tackle most challenging cases except for some extreme cases, it has the capability to serve as an effective auxiliary instrument for diagnosis and surgical planning.

V. Conclusion

In this article, we presented an automatic approach to segment kidneys and tumors from 3-D CT volumes based on an NCG network with scale aware. Unlike conventional 3-D U-Net, we first utilize an SAFE to achieve the adaptive selection of receptive field, which efficiently improves the ability to identify multiscale targets when segmenting kidneys and tumors from 3-D CT volumes. We also propose an NCG mechanism to capture long-range dependencies for feature selections by simultaneously encoding spatial context and recalibrating channel weights. Thanks to the NCG, we can merely use skip connections bridging encoder and decoder in our 3-D U-Net to compensate high-level semantic features with valuable spatial knowledge, achieving more accurate 3-D segmentation. We visually and statistically compare our method with state-of-the-art methods on the KiTS dataset that includes various kidney and tumor cases, demonstrating the advantages and effectiveness of our approach.

References

[1] J. Ferlay et al., “Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods,” *Int. J. Cancer*, vol. 144, no. 8, pp. 1941–1953, 2019.

[2] E. Scosyrev, E. M. Messing, R. Sylvester, S. Campbell, and H. Van Poppel, “Renal function after nephron-sparing surgery versus radical nephrectomy: Results from EORTC randomized trial 30904,” *Eur. Urol.*, vol. 65, no. 2, pp. 372–377, 2014.

[3] U. Capitanio and F. Montorsi, “Renal cancer,” *Lancet*, vol. 387, no. 10021, pp. 894–906, 2016.

[4] T. Heimann and H.-P. Meinzer, “Statistical shape models for 3D medical image segmentation: A review,” *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, 2009.

[5] J. Xie, Y. Jiang, and H.-T. Tsui, “Segmentation of kidney from ultrasound images based on texture and shape priors,” *IEEE Trans. Med. Imag.*, vol. 24, no. 1, pp. 45–57, Jan. 2005.

[6] O. Gloger, K. D. Tönnies, V. Liebscher, B. Kugelmann, R. Laqua, and H. Völzke, “Prior shape level set segmentation on multistep generated probability maps of MR datasets for fully automatic kidney parenchyma volumetry,” *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 312–325, Feb. 2012.

[7] C. Jin et al., “3D fast automatic segmentation of kidney based on modified AAM and random forest,” *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1395–1407, Jun. 2016.

[8] H. Wu, X. Chen, P. Li, and Z. Wen, “Automatic symmetry detection from brain MRI based on a 2-channel convolutional neural network,” *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4464–4475, Sep. 2021.

[9] H. Wu, Z. Zhao, J. Zhong, W. Wang, Z. Wen, and J. Qin, “PolypSeg+: A lightweight context-aware network for real-time polyp segmentation,” *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2610–2621, Apr. 2023.

[10] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[11] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-net: Learning dense volumetric segmentation from sparse annotation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 424–432.

[12] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, “High-resolution encoder–decoder networks for low-contrast medical image segmentation,” *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2019.

[13] K. Kamnitsas et al., “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[14] C. Zhang et al., “HIFUNet: Multi-class segmentation of uterine regions from MR images using global convolutional networks for HIFU surgery planning,” *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3309–3320, Nov. 2020.

[15] S. Feng et al., “CPFNet: Context pyramid fusion network for medical image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[16] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[17] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, “SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network,” *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020.

[18] Z. Wang, N. Zou, D. Shen, and S. Ji, “Non-local U-nets for biomedical image segmentation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6315–6322.

[19] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. van Ginneken, “Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans,” *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2664–2675, Aug. 2020.

[20] Z. Li, J. Pan, H. Wu, Z. Wen, and J. Qin, “Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3D U-net,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 197–206.

[21] J. Li, C. Feng, X. Lin, and X. Qian, “Utilizing GCN and Meta-learning strategy in unsupervised domain adaptation for pancreatic cancer segmentation,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 79–89, Jan. 2022.

[22] X. Yang, Y. Zhang, B. Lo, D. Wu, H. Liao, and Y.-T. Zhang, “DBAN: Adversarial network with multi-scale features for cardiac MRI segmentation,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2018–2028, Jun. 2021.

[23] T. Kitrungrotsakul et al., "Attention-RefNet: Interactive attention refinement network for infected area segmentation of COVID-19," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2363–2373, Jul. 2021.

[24] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019.

[25] Y.-J. Huang et al., "3-D RoI-aware U-net for accurate and efficient colorectal tumor segmentation," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5397–5408, Nov. 2021.

[26] J. Xue et al., "Cascaded MultiTask 3-D fully convolutional networks for pancreas segmentation," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 2153–2165, Apr. 2021.

[27] N. H. Weerasinghe, N. H. Lovell, A. W. Welsh, and G. N. Stevenson, "Multi-parametric fusion of 3D power Doppler ultrasound for fetal kidney segmentation using fully convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2050–2057, Jun. 2021.

[28] L. B. da Cruz et al., "Kidney segmentation from computed tomography images using deep neural network," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103906.

[29] Q. Yu, Y. Shi, J. Sun, Y. Gao, J. Zhu, and Y. Dai, "Crossbar-net: A novel convolutional neural network for kidney tumor segmentation in CT images," *IEEE Trans. Image Process.*, vol. 28, pp. 4060–4074, 2019.

[30] T. Zhang et al., "Semantic attention and scale complementary network for instance segmentation in remote sensing images," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10999–11013, Oct. 2022.

[31] J. Yu, J. Yao, J. Zhang, Z. Yu, and D. Tao, "SPRNet: Single-pixel reconstruction for one-stage instance segmentation," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1731–1742, Apr. 2021.

[32] C. Liu, W. Wang, J. Shen, and L. Shao, "Stereo video object segmentation using stereoscopic foreground trajectories," *IEEE Trans. Cybern.*, vol. 49, no. 10, pp. 3665–3676, Oct. 2019.

[33] X. Du et al., "An integrated deep learning framework for joint segmentation of blood pool and myocardium," *Med. Image Anal.*, vol. 62, May 2020, Art. no. 101685.

[34] S. Graham et al., "MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Med. Image Anal.*, vol. 52, pp. 199–211, Feb. 2019.

[35] X. Wang, X. Jiang, H. Ding, and J. Liu, "Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 3039–3051, 2019.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[37] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3912–3921.

[38] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Med. Image Anal.*, vol. 51, pp. 21–45, Jan. 2019.

[39] X. Li et al., "3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Med. Image Anal.*, vol. 45, pp. 41–54, Apr. 2018.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*.

[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[42] C. Xue et al., "Global guidance network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101989.

[43] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "ConnNet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2518–2529, May 2019.

[44] X. Li, Z. Zhao, and Q. Wang, "ABSSNet: Attention-based spatial segmentation network for traffic scene understanding," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9352–9362, Sep. 2022.

[45] Z. Zheng et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022.

[46] H. Li, Y. Chen, Q. Zhang, and D. Zhao, "BiFNet: Bidirectional fusion network for road segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 8617–8628, Sep. 2022.

[47] Z. Shen, H. Yang, Z. Zhang, and S. Zheng, "Automated kidney tumor segmentation with convolution and transformer network," in *Proc. Int. Challenge Kidney Tumor Segmentation*, Strasbourg, France, 2022, pp. 1–12.

[48] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," 2019, *arXiv:1907.12273*.

[49] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 593–602.

[50] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3759–3768.

[51] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10869–10876.

[52] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, "Non-local spatial propagation network for depth completion," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 120–136.

[53] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5689–5698.

[54] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3516–3525.

[55] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1580–1589.

[56] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8778–8788.

[57] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.

[58] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[59] N. Heller et al., "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, *arXiv:1904.00445*.

[60] P. Bilic et al., "The liver tumor segmentation benchmark (LiTs)," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102680.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[63] L. Yu, X. Yang, H. Chen, J. Qin, and P.-A. Heng, "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 66–72.

[64] O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," in *Proc. Conf. Med. Imag. Deep Learn.*, Apr. 2018, pp. 1–10.

[65] E. Gibson et al., "Automatic multi-organ segmentation on abdominal CT with dense V-networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1822–1834, Aug. 2018.

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

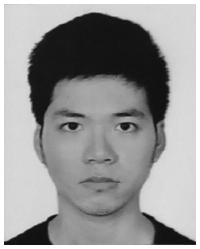
[67] L. Han, Y. Chen, J. Li, B. Zhong, Y. Lei, and M. Sun, "Liver segmentation with 2.5D perpendicular UNets," *Comput. Elect. Eng.*, vol. 91, pp. 107–118, May 2021.

Huisi Wu (Senior Member, IEEE) received the B.E. degree in computer science and the M.E. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from The Chinese University of Hong Kong, Hong Kong, in 2011.

He is currently a Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests are in computer graphics, image processing, and medical imaging.

Baiming Zhang received the B.S. degree in computer science and technology from Shenzhen University, Shenzhen, China, in 2020, where he is currently pursuing the master's degree.

His research interests include computer vision, medical image segmentation, and pattern recognition.



Zhuoying Li received the B.S. degree in software engineering from the Changsha University of Science and Technology, Changsha, China, in 2018. He is currently pursuing the master's degree with Shenzhen University, Shenzhen, China.

His research interests include computer vision, medical image segmentation, and pattern recognition.

Jing Qin (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2009.

He is currently an Associate Professor with the School of Nursing, Hong Kong Polytechnic University, Hong Kong, where he is also a Key Member with the Centre for Smart Health. He has participated in over 10 research projects and published over 90 papers in major journals and conferences in the below areas. His current research interests include virtual/augmented reality for healthcare and medicine training, medical image processing, deep learning, visualization and human–computer interaction, and health informatics.

Tong-Yee Lee (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Washington State University, Pullman, WA, USA, in May 1995.

He is currently a Chair Professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University (http://graphics.csie.ncku.edu.tw). His current research interests include computer graphics, nonphotorealistic rendering, medical visualization, virtual reality, and media resizing.

Dr. Lee is a Member of the ACM.