BMC Medical Imaging

RESEARCH Open Access

Enhanced Cross-stage-attention U-Net for esophageal target volume segmentation

Xiao Lou[1,2], Juan Zhu[4], Jian Yang[5], Youzhe Zhu[1,2*], Huazhong Shu[1*] and Baosheng Li[1,3*]

Abstract

Purpose The segmentation of target volume and organs at risk (OAR) was a significant part of radiotherapy. Specifically, determining the location and scale of the esophagus in simulated computed tomography images was difficult and time-consuming primarily due to its complex structure and low contrast with the surrounding tissues. In this study, an Enhanced Cross-stage-attention U-Net was proposed to solve the segmentation problem for the esophageal gross tumor volume (GTV) and clinical tumor volume (CTV) in CT images.

Methods First, a module based on principal component analysis theory was constructed to pre-extract the features of the input image. Then, a cross-stage based feature fusion model was designed to replace the skip concatenation of original UNet, which was composed of Wide Range Attention unit, Small-kernel Local Attention unit, and Inverted Bottleneck unit. WRA was employed to capture global attention, whose large convolution kernel was further decomposed to simplify the calculation. SLA was used to complement the local attention to WRA. IBN was structed to fuse the extracted features, where a global frequency response layer was built to redistribute the frequency response of the fused feature maps.

Results The proposed method was compared with relevant published esophageal segmentation methods. The prediction of the proposed network was MSD = 2.83(1.62, 4.76)mm, HD = 11.79 ± 6.02 mm, DC = 72.45 ± 19.18% in GTV; MSD = 5.26(2.18, 8.82)mm, HD = 16.22 ± 10.01 mm, DC = 71.06 ± 17.72% in CTV.

Conclusion The reconstruction of the skip concatenation in UNet showed an improvement of performance for esophageal segmentation. The results showed the proposed network had better effect on esophageal GTV and CTV segmentation.

Keywords Esophageal carcinoma, Simulated CT, Esophageal segmentation, CNN, UNet, Attention

*Correspondence:
Youzhe Zhu
zhuyouzhe@yeah.net
Huazhong Shu
shu.list@seu.edu.cn
Baosheng Li
bsli@sdfmu.edu.cn
[1] Laboratory of Image Science and Technology, Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Southeast University, Sipailou 2, Nanjing, P.R. China
[2] Department of Radiotherapy, Lishui People's Hospital, No. 1188, Liyang Street, Lishui, P.R. China
[3] Shandong First Medical University and Shandong Academy of Medical Sciences, Shandong Cancer Hospital and Institute, No. 440, Jiyan Street, Jinan, P.R. China
[4] Department of Respiratory Medicine, The People's Hospital of Zhangqiuqu Area, No. 1920, Huiquan Street, Jinan, P.R. China
[5] Department of Clinical Laboratory, The People's Hospital of Zhangqiuqu Area, No. 1920, Huiquan Street, Jinan, P.R. China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

In the radiotherapy processes for cancer, developing the radiotherapy plan was one of the most important steps. As a prerequisite for making an esophageal radiotherapy plan, automatic segmenting technique for the target volume played a decisive role [1–3]. This could enable radiation oncologists to lock the location and shape of the lesions promptly, thereby assisting them in determining the radiation range and delineating the target volume [4]. In addition, due to the complex structure and random position of the esophagus, as well as its lower contrast with surrounding organs and tissues [5, 6], manually outlining the esophageal gross target volume (GTV) and clinical target volume (CTV) was time-consuming and laborious. Esophageal segmentation technique could help radiation oncologists initialize the range of the target volume, greatly improving the efficiency for making the radiotherapy plan [7–10]. Moreover, after delineating the target area, esophageal image segmentation technique can help medical physicists distribute the range and dose of the radiation, thus to reduce the exposure of organs at risk (OAR) when making specific radiotherapy plans. In summary, the objective of esophageal segmentation technique was to ensure the adequate dose of radiation to the lesion while keeping the dose of OAR as low as possible.

In 2006, Rousson et al. proposed a method for extracting the most probability of extracting the esophagus from the Computed Tomography (CT) image which used the probability space model to extract the central line of the esophagus [11]. In the same year, Huang et al. proposed a semi -automatic esophageal segmentation method based on a certain CT slice which predetermined the esophageal segmentation, and the outline was used to spread the outline to other slices [12]. Fieselmann et al. proposed an esophageal segmentation method in 2008, which converted the esophageal contour to the frequency domain for segmentation [13]. Feulner et al. proposed a method based on "detection and connection" to estimate the approximate shape of the esophagus, and classified the subtle section with a classifier trained by elliptical models [14]. Kurugol et al. proposed a method in 2011, using the organization around the esophagus to determine the position of the esophageal center line in a specific CT slice, and then using level set method to extract the esophageal contour [15]. Yang et al. proposed a multi -map segmentation method in 2017, selecting the best maps of the spectrum for esophageal segmentation [16].

Nowadays, because of the intelligence and high precision, deep learning method was widely explored in image segmentation area. You et al. proposed an approach based on the Wasserstein distance guided disentangled representation to achieve 3D multi-domain liver segmentation in 2020 [17]. They proposed a novel type of adversarial transformers named CASTformer for 2D medical image segmentation in 2022 [18]. They proposed a novel multi-site segmentation framework named incremental-transfer learning (ITL), which learned a model from multi-site datasets in an end-to-end sequential fashion in 2022 [19]. They proposed a simple contrastive distillation framework named SimCVD that significantly advanced state-of-the-art voxel-wise representation learning in 2022 [20]. They presented a contrastive voxel-wise representation learning (CVRL) method to effectively learn low-level and high-level features by capturing 3D spatial context and rich anatomical information along both the feature and the batch dimensions in 2022 [21]. They proposed an anatomical-aware contrastive distillation framework named ACTION for semi-supervised medical image segmentation in 2023 [22]. They proposed an improved contrastive learning framework named ACTION+ +with adaptive anatomical contrast for semi-supervised medical segmentation in 2023 [23]. They proposed a generic implicit neural rendering framework named MORSE designed at an anatomical level to assist learning in medical image segmentation in 2023 [24]. They proposed a novel semi-supervised 2D medical image segmentation framework termed Mine your own anatomy (MONA) in 2024 [25]. They proposed a semi-supervised contrastive learning (CL) framework named ARCD with stratified group theory for medical image segmentation in 2024 [26]. Zhu et al. proposed a 3D end-to-end fully convolutional neural network named semantic V-net (SV-net) for segmentation in 2021 [27]. Liu et al. proposed a network composed of transformer and convolution branches to fuse the global and local information in 2023 [28]. They proposed a novel weakly supervised biomarkers localization and segmentation method named TSSK-Net which required only image-level annotations in 2023 [29]. Yang et al. proposed a novel Discrete Wavelet Transform and Attention (DWTA) module to precisely predict the crack region based on the famous skeleton network Unet in 2022 [30]. Geng et al. proposed a novel coronavirus image segmentation network alternately using Swin transformer and CNN (STCNet) in 2024 [31].

In recent years, due to the excellent performance of neural networks, deep learning methods have gradually been used in the field of esophageal segmentation in medical images. In 2017, Hao et al. proposed an esophageal segmentation method based on FCN and graph -cut technique [32]. Trullo et al. modified the FCN structure to improve the accuracy of the esophageal positioning and segmentation in the same year [33]. They developed the Sharpmask (SM) architecture to better connect the features of deep and shallow layers. In 2019, Chen et al. used the U-Net Plus network to segment the esophagus in CT images [34]. In 2020, Diniz et al. used an ATLAS based deep learning method to pre-segment the esophagus as a data set for training neural networks [35]. Yousefi et al. proposed a sparse dense attention UNet (DDAUNET) in 2021, which used spatial and channel attention gates in the dense module [36]. In the same year, Tan et al. proposed an esophageal segmentation of the neural network based on attention mechanisms and space pyramid modules (SAN) [37]. Alam et al. proposed a 3D-UNet which divided the scope of the simulated CT and CBCT respectively in 2021 as well [38]. Jin et al. proposed a progressive semantical-network (PSNN) in the same year, segmenting the esophagus in simulated CT images [39]. Li et al. proposed a hybrid attention network (HAN) to address the esophageal segmentation problem in Optical Coherence Tomography (OCT) images in 2024 [40]. Jian et al. proposed an architecture named HRU-Net for esophageal cancer and esophageal carcinoma segmentation in CT slices in 2024 [41].

The previous esophageal segmentation works were mainly divided into traditional segmentation methods and deep learning methods. Traditional methods were mainly composed of semi-automatic methods which usually needed to pre-mark the key point of the organs or directly pre-segment the organs surrounding the esophagus. In general, the traditional methods required a lot of priority knowledge, and the performance was sensitive to the modulation of parameters. For deep learning methods, Vision Transformer (ViT) was a widely employed framework in medical image segmentation, which focused on the global features of the images and structured the long dependence among the image contents. However, it usually required a large scale of parameters as support and lacked of local information. The segmentation accuracy of this framework was unstable as well when the sample size was small. Compared with ViT, the UNet network, which based on convolutional neural network (CNN), had a compact structure and was suitable for multi-scale tasks, especially for medical image segmentation tasks with small sample size. The majority of existing deep learning based esophageal segmentation methods are modified on the foundation of UNet, such as Sharpmask [33], Res-UNET [35], DDAUNET [36], SAN [37], PSNN [38]. However, the features of different depths in UNet had disparate degrees of abstraction. However, these networks did not pay close attention to the extraction of global features, which was the main advantages of ViT. Moreover, because each stage of the encoder in UNet owned different levels of features, the direct concatenation between the deep and shallow layers had the risk of destroying the structure of features. In addition, plenty of segmentation methods did not preprocess the input of the network, which was to the benefit of model training. Based on this research background and the clinical requirements, this study proposed an Enhanced Cross-stage-attention U-Net (ECAU), which aimed to reconstruct the skip concatenation in UNet and promote the esophageal segmentation of GTV and CTV in simulated CT images, to pre-extract the image features of the input, and to capture the global attention among the fusion-stage feature maps.

Method

ECAU constructed a series of CNN based fusion module——Enhanced Cross-stage-attention Block (ECB) to replace the skip concatenation in U-Net, and the structure of ECAU was shown in Fig. 1. First, the initial features were extracted and purified by the Principal Channels Extraction (PCE) module. Then the feature maps were transferred to main network with UNet type backbone. For the fundamental units in ECB, the Wide Range Attention unit (WRA) was designed

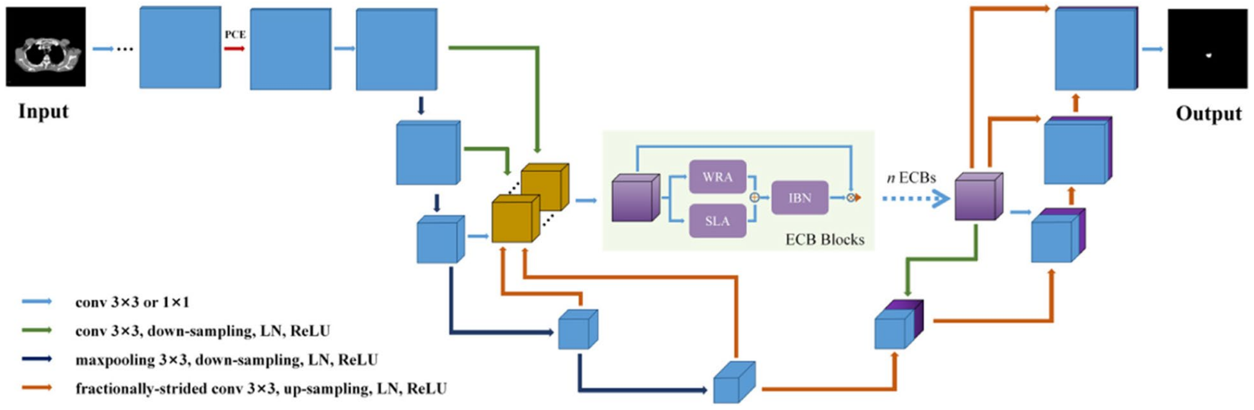

Input
PCE
Output
WRA
SLA
IBN
ECB Blocks
n ECBs
conv 3×3 or 1×1
conv 3×3, down-sampling, LN, ReLU
maxpooling 3×3, down-sampling, LN, ReLU
fractionally-strided conv 3×3, up-sampling, LN, ReLU

Fig. 1 Structure of the proposed neural network ECAU

to expand the receptive field, whose depth-wise convolution layer was further decomposed to reduce the computation; the Small-kernel Local Attention unit (SLA) was structured to enhance the capture of local attention, as a complement to WRA; subsequently, the Frequency Global Response Normalization Layer (FGRNL) was constructed in the Inverted Bottleneck unit (IBN) after the feature fusion of WRA and SLA, which strengthened the comparability and selectivity of the feature maps among the channels and improved the competitiveness of "important" features in frequent domain. Because this study primarily emphasis on the substitute of the skip concatenation in UNet, the original UNet was employed as the backbone of the proposed neural network.

The structure of PCE module

In general, the channels would be expanded after the image inputting to the network to achieve richer feature information. However, there might be a certain degree of correlation among the expanded channels which leaded to a redundancy. The PCE module was designed in the channel expansion stage of the neural network as an initialization of feature map. The objective of PCE was to extracted the principal features in channel dimension while eliminating the redundant feature maps in the meanwhile. The workflow of PCE was shown in Fig. 2.

In PCE, the theory of principal component analysis [42] was employed to map the feature vectors in channel dimension to orthogonal vectors, in order to reflect the information as much as possible by using less quantity of channels. The flow chart of PCE was shown in Fig. 3. H and W were the height and width of the feature maps, respectively, and C represented the quantity of channels. The primary processes of PCE were: First, the feature maps were unfolded in $H\times W$ dimension, and a two-dimensional (2D) matrix was constructed by $H\times W$ samples with C variates (the matrix needed to be transferred); Second, the PCA method was employed to simplify the quantity of features using the method of singular value decomposition (SVD); Finally, the feature matrix was reshaped to the three-dimensional (3D) form, consistent with the input feature maps (the dimension of batches was ignored). Apart from extracting the primary features, the PCE module performed similar to the element-wise convolution layer, which had the capacity to reduce the channels into the required quantity. This module could be regarded as a preliminary collection and filtration of image features.

The structure of ECB module

In the neural network, a large receptive filed generally had a powerful capability for information extraction. Some studies have proved that the global attention mechanism of ViT can be simulated by large depth-wise kernel in CNN to achieve a similar receptive field, which could accelerate the calculation in the meanwhile [43, 44]. However, large convolution kernel of CNN still consumed great computing resource of GPU, especially occupying the memory of graphic card. In ECB, the WRA module was explored to enhance the scale of the depth-wise convolution kernel, and to further split the convolution kernel to several sublines in parallel. The SLA employed small depth-wise convolution kernel to capture the local attention, as a supplement to WRA. Subsequently, the IBN was structured to fuse the attentions and reconstruct the frequency response of the features. The specific structure of ECB was shown in Fig. 4.

The structure and function of WRA unit

In this study, the goal size of depth-wise convolution kernel in WRA was 28×28, which was split into a bunch of parallel layers. First, the layers were transformed to a general convolution layer with the kernel_size = 7×7, and an atrous convolution layer with the kernel_size = 6×6, dilation = 4, and padding = 10. However, the convolution layer with kernel_size = 7×7 still had burden on GPU. According to statistics, the consumption of the convolution with kernel_size = 7×7 was about 1.4 times the kernel_size = 3×3 under the same floating-point calculation conditions [45]. In addition, some work demonstrated that not all of the channels in the depth-wise convolution layer had the same importance [46]. In this study, the depth-wise convolutional layer with kernel_size = 7×7 was decomposed to 4 independent parallel units, which was shown in Eq. (1):

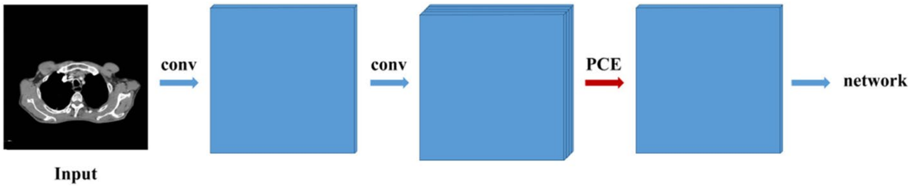


Fig. 2 Primary features pre-extraction by PCE module

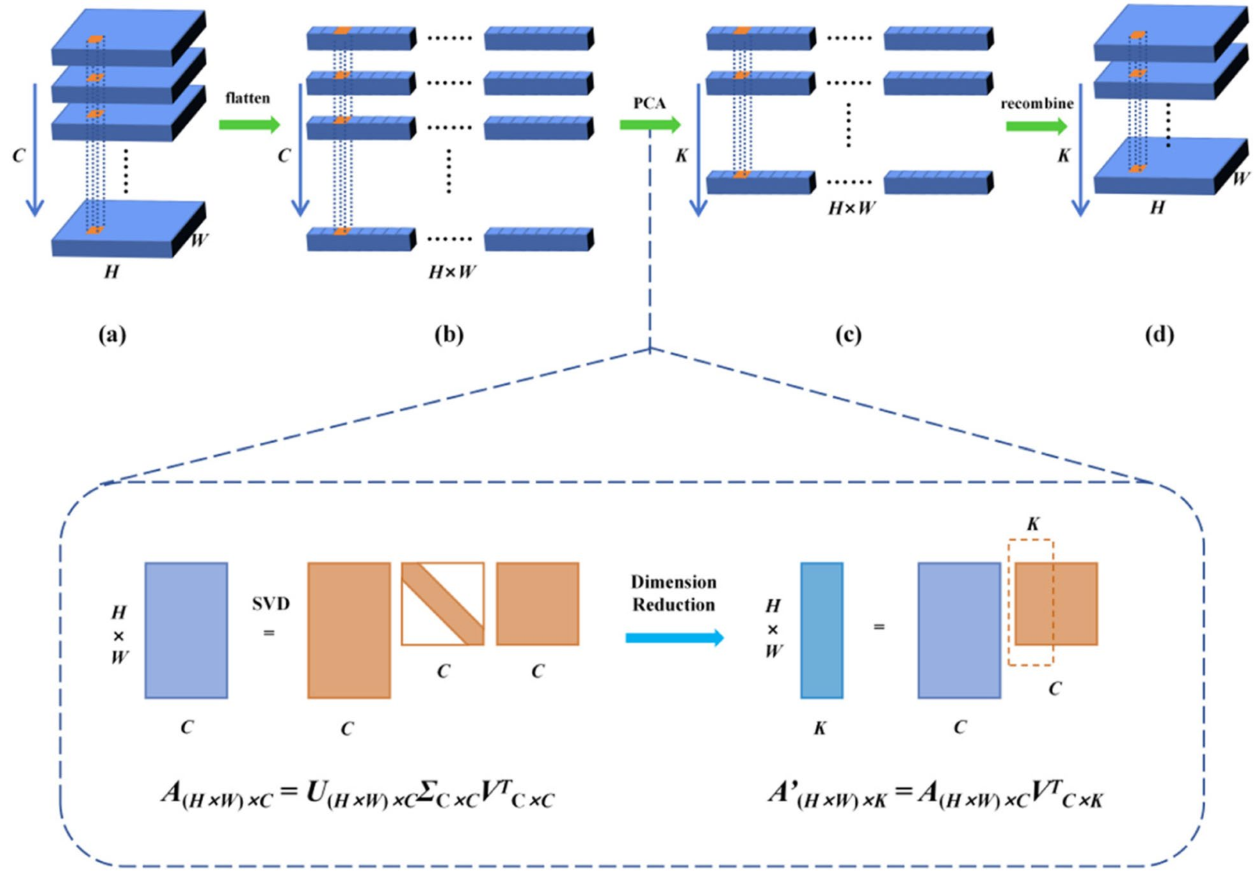

C
flatten
C
H×W
PCA
K
H×W
recombine
K
H
W
H
W
(a)
(b)
(c)
(d)
H × W
C
SVD =
C
C
C
Dimension Reduction
H × W
K
=
C
K
C
$A_{(H\times W)\times C} = U_{(H\times W)\times C}\Sigma_{C\times C}V^{T}_{C\times C}$
$A'_{(H\times W)\times K} = A_{(H\times W)\times C}V^{T}_{C\times K}$

Fig. 3 Mechanism and flow chart of PCE

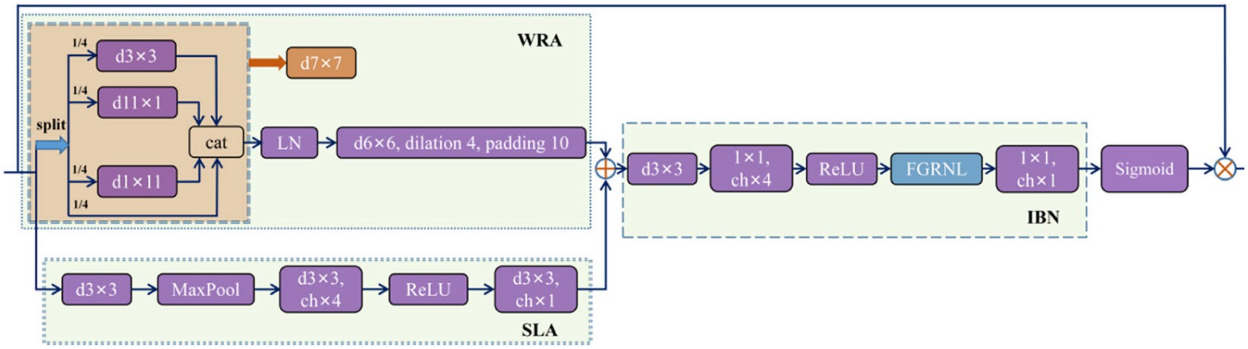


Fig. 4 Structure of ECB module

$$\mathrm{Conv}_{\mathrm{dw},7\times7} \approx \mathrm{Concat}(X_i)\ X_i = \begin{cases} \mathrm{Conv}_{\mathrm{dw},3\times3}\left[\frac{1}{4}\mathrm{Split}(\mathrm{WRA}_{\mathrm{in}})\right], & i = 1 \\ \mathrm{Conv}_{\mathrm{dw},11\times1}\left[\frac{1}{4}\mathrm{Split}(\mathrm{WRA}_{\mathrm{in}})\right], & i = 2 \\ \mathrm{Conv}_{\mathrm{dw},1\times11}\left[\frac{1}{4}\mathrm{Split}(\mathrm{WRA}_{\mathrm{in}})\right], & i = 3 \\ \frac{1}{4}\mathrm{Split}(\mathrm{WRA}_{\mathrm{in}}), & i = 4 \end{cases} \tag{1}$$

where $\mathrm{WRA}_{\mathrm{in}}$ represented the input of WRA unit, Split represented the separation of this input in proportion, $\mathrm{Conv}_{\mathrm{dw}}$ represented the depth-wise convolution, X_1 branch represented the compressed convolution, X_2 and X_3 branches ensured the receptive field after the decomposition of the large kernel, and X_4 retained a part of the original channel. The effectiveness of splitting the large kernel convolution layer to $K\times1$ and $1\times K$ convolutional layer was confirmed by InceptionV3 [47]. This split was in fact a form of sparse convolution. In addition, not exactly the same as InceptionNeXt, the large kernel was not merely split into three small kernels with a ratio of 1/3 in this study, but reserving an unprocessed channel individually. This operation was in order to avoid the problem of gradient disappearance during the training process, which referred to the mapping design in the residual network.

The structure and function of SLA unit

The role of SLA unit was primarily to complement the function of WRA in terms of attention capture. In theory, as long as expanding the quantity of channels in WRA, the information content in the convolution layer with large kernel size could cover that with small kernel size. However, to capture the local attention in CNN pattern by purely augmenting the channels in WRA would exaggeratively increase the parameters and computation quantity during training. Therefore, the SLA was aggrandized parallelly with WRA to enhance the capture capability for local attention.

The IBN unit with FGRNL module

After the WRA and SLA units capturing the global and local attentions, these features were fused by IBN. First, the channels were stacked and expanded to enrich the feature maps, but it would cause unavoidable feature redundancy [48]. The way to filter and exclude extra information could be found in biology, such as lateral inhibition in neurons [49]. Dynamic feature door control method [50–52] had similar effects. For example, Squeeze-and-excite (SE) method [53] employed the space gate control; Convolutional Block Attention Module (CBAM) employed channel gate control [54]. However, these methods had an excess of parameters to burden the calculation during training. Recently, there was a method named Global Response Normalization (GRN), using the L2 norm as the weight to evaluate the importance of each channel [48]. However, this method only used the "energy" information in each channel, which was not able to cover the valuable information entirely. For example, the proportion of energy in edges was slight, but it contained more considerable features than smooth parts.

According to Parseval's theorem, the energy (L2 norm) in space domain was conserved with the energy in frequency domain, whose discrete form was shown in Eq. (2):

$$\sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |X(k)|^2 \tag{2}$$

where $x(n)$ is the pixel value in the spatial domain, $X(k)$ is the pixel value in the frequency domain, and N is the total number of pixels. In this study, the FGRNL module was designed after the channel expansion layers in IBN, which measured the L1 norm as the weight of each frequent component in the channel dimension. The FGRNL enhanced the selectivity of each frequent component in frequent domain, and the processes were shown in Fig. 5, which was mainly divided into four steps:

(1) Spatial domain transformation. The feature maps were transformed to frequent domain in $H\times W$ dimension, as shown in Eq. (3):

$$F^C(u,v) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f^C(x,y)e^{-j2\pi(ux/M+vy/N)}$$
$$u = 0,1,2,\ldots,M-1;\ v = 0,1,2,\ldots,N-1 \tag{3}$$

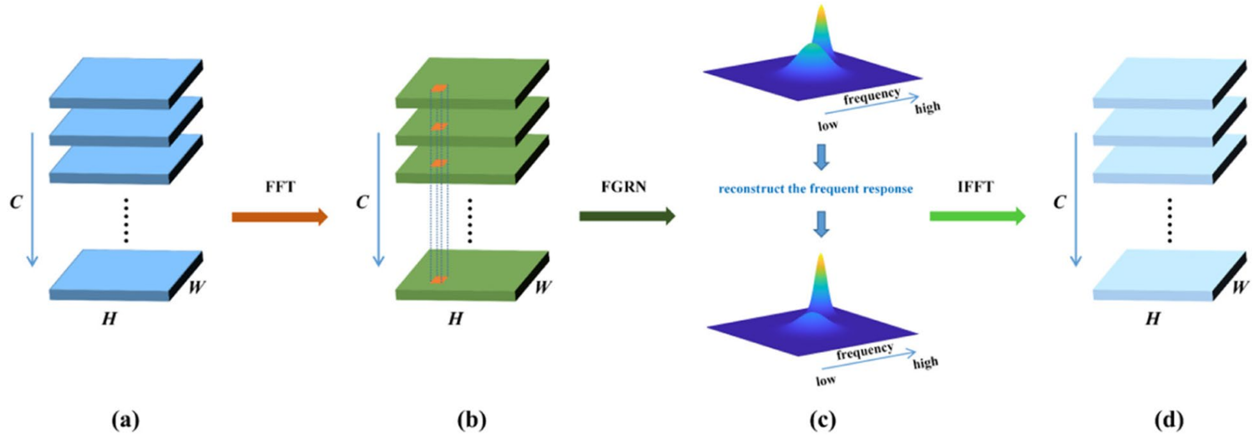


Fig. 5 Mechanism and flow chart of FGRNL module

where, x and y were the variables of the spatial domain in H and W dimensions, respectively; u and v were the corresponding variables in the frequency domain of the feature maps; M and N were the number of pixels in the corresponding dimensions; C represented the channel dimension; $f^C(x,y)$ was the feature map of each channel and $F^C(u,v)$ was its discrete spectrum.

(2) Frequent variable standardization. The amplitudes of the frequent components were calculated firstly. Then, the L1 norm of each component in channel dimension was counted. The processes were shown in Eq. (4) and Eq. (5):

$$\Gamma^C(u,v) = \left(\left|F^1(u,v)\right|, \left|F^2(u,v)\right|, ..., \left|F^C(u,v)\right|\right) \\ u = 0, 1, 2, ..., M-1;\ v = 0, 1, 2, ..., N-1 \tag{4}$$

$$\left\|\Gamma^C(u,v)\right\|_1 = \sum_{i=1}^{C} \left|F^i(u,v)\right| u = 0, 1, 2, ..., M-1;\ v = 0, 1, 2, ..., N-1 \tag{5}$$

where $\Gamma^C(u,v)$ was the modulus calculated in the frequent domain, and $\left\|\Gamma^C(u,v)\right\|_1$ was its L1 norm.

(3) Global frequency response evaluation. The frequent amplitudes of all channels were normalized, as shown in Eq. (6):

$$r^C(u,v) = \Gamma^C(u,v) \Big/ \left\|\Gamma^C(u,v)\right\|_1 u = 0, 1, 2, ..., M-1;\ v = 0, 1, 2, ..., N-1 \tag{6}$$

where $r^C(u,v)$ was the frequent normalized ratio of each channel. This step inhibited the unconsidered frequency in each channel, emphasized the significance of the primary frequency, and strengthened the frequent competition among channels. This ratio was subsequently multiplied with the original spectrum (array multiplication) to calibrate the frequency response of each channel, as shown in Eq. (7):

$$F^C_{\text{calibrated}}(u,v) = F^C(u,v) * r^C(u,v)\ u = 0, 1, 2, ..., M-1;\ v = 0, 1, 2, ..., N-1 \tag{7}$$

where * was the symbol of array multiplication and $F^C_{\text{calibrated}}(u,v)$ represented the spectrum of the feature maps after calibration.

Frequent domain inverse transformation. The calibrated feature maps were converted back to the spatial domain by discrete Fourier inversion method, as shown in Eq. (8):

$$f^C(x,y) = \sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F^C(u,v) e^{j2\pi(ux/M+vy/N)}\ x = 0, 1, 2, ..., M-1;\ y = 0, 1, 2, ..., N-1 \tag{8}$$

In order to facilitate the training process, two trainable parameters α and β were added to the FGRNL module, as shown in Eq. (9):

$$f^C_{\text{final}}(x,y) = \alpha \times f^C(x,y) + \beta\, x = 0, 1, 2, ..., M-1;\ y = 0, 1, 2, ..., N-1 \tag{9}$$

where the scales of α and β were the same as $f^C(x,y)$.

Data set and experimental parameters

Data set of the experiment

Study population

This study included 124 patients with esophageal carcinoma and the distribution of carcinoma in the population was shown in Table 1. The lesions primarily concentrated in thoracic parts which was the main part of the esophagus. The majority of tumor staging exceeded T2, N1 in TNM

Table 1 Distribution of population with esophageal carcinoma in this study

Tumor Classification			Number of Patients
Tumor Location		Cervical part	16
		Upper thoracic part	36
		Middle thoracic part	40
		Lower thoracic part	32
TNM Staging	T Staging	T1	6
		T2	16
		T3	91
		T4	11
	N Staging	N0	12
		N1	28
		N2	60
		N3	24
	M Staging	M0	102
		M1	22
Clinical Staging		I	4
		II	12
		III	75
		IV	33
Pathology Type		Squamous carcinoma	114
		Small cell carcinoma	7
		Neuroendocrine carcinoma	3

Staging and II in Clinical Staging, respectively. The pathology type of the esophageal carcinoma in this study mainly consisted of squamous carcinoma. The distribution of the entire data set basically accorded with the disease type of the local patients.

Data acquisition

All of the patients achieved normal treatment, but not for clinical trials or scientific research. The CT scans of the patients entirely contained 19,716 simulated CT images. The simulated CT scan generally involved the range of the entire esophagus, but the target volume only occupied a portion of the slices that contained the lesion. Therefore, the slices without lesions were excluded in this study. As a result, a total of 4985 simulated CT images comprised the data set after filtration. The images were provided by the Department of Radiotherapy in Lishui People's Hospital and scanned by Siemens Somatom Definition AS + simulator. The X-ray level of the simulator was 120 kV, the window level was 50, the window width was 350, the DFOV was 450 mm, and the thickness of the slice was 3–5 mm. The scale of the images was 512 × 512 pixels, and the corresponding spatial scale was $450 \times 450\text{mm}^2$.

Segmentation

All of the esophageal GTV and CTV were delineated by a radiation oncologist and verified by other two radiation oncologists, based on MIM system (MIM software Inc.). These radiation oncologists had similar seniority and experience in clinical practice. If there was a division of opinion on this issue, they would have a discussion according to the comprehensive materials of the patients and reached an agreement in most cases. Otherwise, the contention would be presented in the department-wide meeting and abode by the majority opinion. After delineation, the CT images and tumor volumes would be transformed from DICOM to common image file format (png). The pixel values in CT images were truncated according to the window level and window width, and normalized in the reconstructed range. The delineated tumor volumes were transformed from RTstruct to binary images and the internal regions contained in the target contours were the ground truth.

Parameters testing of ECB

After repeatedly trial, the quantity of channels importing the ECBs was assigned to 256 and the skip concatenation of UNet was replaced by 4 ECBs to balance the performance and computing resource. The output of ECBs was re-scaled by 3 × 3 convolution and fractionally-strided convolution in several branches to match the scale of the feature maps in up-sampling stages of ECAU.

Devices and hyper-parameters of neural network

The image data sets were divided into training, validation, and test sets with a ratio of 3:1:1 and augmented by cropping, rotation, space zooming and horizontal flipping before training. The core hardware of the experiment was NVIDIA GeForce GTX 3070 Ti graphics card. The neural networks were programmed based on Python 3.7 and PyTorch framework in PyCharm Integrated Development Environment (IDE). In terms of neural network hyperparameters, the Adam optimizer was used for network optimization. The training epoch was 300 with the batch size of 4. The combination of binary focal loss and Dice loss was the loss function with the ratio of 1:1. Dice coefficient (DC) was screened accuracy. ReducelronPlateau was the function for learning rate adjustment with the initial learning rate of 0.001, where the learning rate would be multiplied by 0.5 if the screened accuracy was not optimized in 10 epochs except for the initial 5 epochs during training.

Results and discussion

Performance of PCE and FGRNL modules

In order to measure the performance of PCE and FGRNL module, an ablation experiment was operated, and the DC was used to evaluate the prediction, which was shown in Table 2. The prediction accuracy in these situations showed a stepped-up trend, which was more prominent in CTV than that in GTV. This was primarily because for the ROI (range of interest) of GTV, the boundary was closer to the periphery of the esophagus than that in CTV, and the contrast between exterior and interior was more distinct, so the features were easier to extract. In comparison, determining the ROI of CTV was more dependent on the clinical experience of the radiation oncologist, which had lower correlation to the shallow features such as edges. Therefore, although the

Table 2 Ablation experiment for PCE and FGRNL module (DC)

	Simple ECAU (%)	With PCE (%)	With FGRNL (%)	With PCE & FGRNL (%)
GTV	71.05 ± 19.50	71.93 ± 16.80	72.33 ± 18.37	72.45 ± 19.18
CTV	64.12 ± 17.14	67.25 ± 16.70	69.14 ± 16.72	71.06 ± 17.72

esophageal GTV had a more restricted scale than CTV, the overall prediction accuracy of GTV supported by evident features was higher than that of CTV. An example of predicted images of the ablation experiment was shown in Fig. 6.

Comparison of accuracy among the networks

In order to further evaluate the segmentation capacity of ECAU, it was compared with the relevant esophageal segmentation methods based on the same data set provided by this study. The evaluation criteria were mean surface distance (MSD), Hausdorff distance (HD), and DC. MSD measured the average match between two graphs by calculating the distance of each pixel in one graph from the other, and then calculating the average of entire pixels. Supposing there were two contours in the image, and the pixels of the contours were represented by pixel sets $A=(a_1, a_2,..., a_p)$ and $B=(b_1, b_2,..., b_q)$, then MSD could be expressed as Eq. (10):

$$M(A,B) = \max\left[m(A,B), m(B,A)\right] \tag{10}$$

where

$$m(A,B) = (1/p)\sum_{i=1}^{p} \min_{b\in B}\left\|a_i - b\right\| \tag{11}$$

$$m(B,A) = (1/q)\sum_{i=1}^{q} \min_{a\in A}\left\|b_i - a\right\| \tag{12}$$

Equation (11) and Eq. (12) represented the average distance from all points in point set A to another point set B, and vice verse. $\min_{b\in B}\left\|a_i - b\right\|$ represented the minimum Euclidean distance from point a_i in pixel set A to pixel set B, and p represented the number of pixels in pixel set A. The definition of Eq. (12) was similar to Eq. (11). Equation (10) indicated that MSD ultimately adopted the maximum of $m(A, B)$ and $m(B, A)$.

In contrast to MSD, HD measured the farthest distance between two graphs. The basic distance calculation method of pixels from one graph to another in HD was similar to that of MSD. The difference was that HD utilized the maximum value of the corresponding distance of the pixel set,

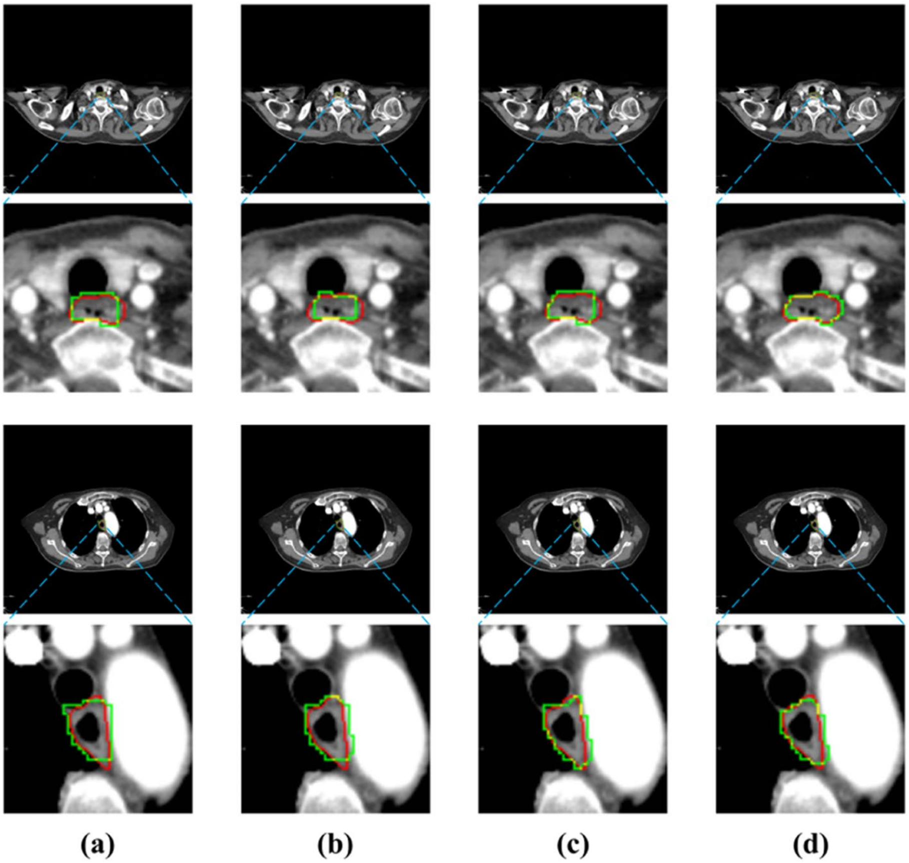


Fig. 6 Mechanism and flow chart of FGRNL module. (**a**) represented the simple ECAU; (**b**) represented ECAU with PCE; (**b**) represented ECAU with FGRNL; (**c**) represented ECAU with PCE & FGRNL

rather than the average value, in order to measure the maximum mismatch between the two point sets. HD could be defined as Eq. (13):

$$H(A,B) = \max\left(h(A,B), h(B,A)\right) \tag{13}$$

where

$$h(A,B) = \max_{a \in A} \min_{b \in B} \left\| a_i - b \right\| \tag{14}$$

$$h(B,A) = \max_{b \in B} \min_{a \in A} \left\| b_i - a \right\| \tag{15}$$

$\min_{b \in B} \left\| a_i - b \right\|$ represented the minimum Euclidean distance from point a_i in pixel set A to pixel set B. Equation (14) represented the maximum of the distances of each pixel in point set A. The definition of Eq. (15) was similar to Eq. (14). Equation (13) represented HD would adopt the maximum of $h(A,B)$ and $h(B,A)$.

DC was a measure of similarity between two graphics. In medical image segmentation, DC was frequently used to compare two certain areas. In this study, DC was employed to measure the similarity of the internal areas of the esophagus, rather than the contours. Supposing there were two regions defined as set X and set Y, then DC could be defined as Eq. (16):

$$DC = (2|X \cap Y|)/(|X| + |Y|) \tag{16}$$

The predictions of GTV and CTV were measured respectively, which was shown in Table 3. Due to the inhomogeneity of the esophageal lesion in CT slices, the MSD of the data presented a skewed distribution. Therefore, the skewness coefficient was employed to measure the degree of data skewness, and the median and quartile were used to measure the average and variation levels of the data. Other indicators approximated normal distributed, which was evaluated by mean and standard deviation.

In Table 3, SKEW represented the skewness coefficient, M represented the median, P_{25} represented the lower quartile, P_{75} represented the upper quartile, AVG represented the arithmetic mean, and SD represented the standard deviation. The proposed segmentation method had a decent performance in majority of indicators. For instance, the DC was 72.45 ± 19.18% and 71.06 ± 17.72% respectively in GTV and CTV, which was both superior to other methods. For ECAU, the skewness coefficient of MSD was SKEW = 2.81 and SKEW = 2.76 in GTV and CTV, respectively, which was at a lower level than other methods, indicating that its prediction in different slices of the esophageal lesion was more balanced. Predictive stability was a considerable factor to evaluate the performance of the neural network. Although the median, quartile and HD mean values of DDAUNet in GTV were slightly better than ECAU, its prediction accuracy was instable due to the high skewness coefficient, which was also reflected in its standard deviation of HD. Because the MSD and HD focused on estimating the boundaries, they were easier affected by the abnormal outline of the prediction, which lead to the unconspicuous advantage of the proposed network. In CTV, the proposed network had the best MSD = 5.26(2.18, 8.82)mm and HD = 16.22 ± 10.01 mm, respectively, which also gained advantages in GTV.

The prediction accuracy of most neural networks in GTV was higher than that in CTV, which was consistent with the subjective assessment in Fig. 6. The performances of SharpMask and SAN which had the simpler structure were distinctly disparate in GTV and CTV, which probably due to the restriction of receptive field and the curtness of feature fusion.

Figure 7 showed the examples of the predictions listed in Table 3, where the order of the images from top to bottom corresponding to the vertical axes of anatomy. Subjectively to evaluate the images in Fig. 7, the prediction in GTV was more accurate than that in CTV, which was consistent with the data in Table 3. The predictions in upper thoracic part and lower thoracic part of the esophagus had a higher precision than that in middle thoracic part of the esophagus, especially of the SharpMask and SAN network. The prediction of ECAU approximated to the annular distribution, according with the definition of GTV and CTV, where the local false positive (FP) regions in "branch" type were fewer than other networks. It was observed that the prediction of DDAUNet was most close to the proposed network, followed by Residual UNet, PSNN, SharpMask, and SAN in sequence, which was consistent with the objective evaluation in Table 3.

Figure 8 showed the Bland–Altman graphs which could reflected the matching degree between the predictions and the ground truth. Because approximate normal distribution of the data was a requisite to make the Bland–Altman analysis, the measurements of HD and DC were presented in Fig. 8 except for MSD. The cyan and purple points in the graphs represented the cases in test sets of GTV and CTV, respectively. The red and blue transverse lines represented the upper and lower bounds of limits of agreement (LoA), respectively, which indicated the limit of agreement with the 95% confidence interval. Smaller region between the bounds of LoA meant the network was more reliable. The proposed network had the DC LoA region of [11.88, 26.48] in GTV and [11.15, 24.48] in CTV, and the HD LoA region of [3.03, 9.01] in GTV and [4.76, 15.26] in CTV, which had an advantage to other networks. Moreover, the aggregation degree of the points of proposed network represented the stability of the prediction.

Table 3 comparison between the proposed method and published esophageal segmentation methods (the optimal result was in bold)

Network Metrics			Proposed network	SharpMask [33]	Residual UNet [35]	DDAUNet [36]	SAN [37]	PSNN [38]	HAN [40]	HRU [41]
GTV	MSD (mm)	SKEW	2.81	7.91	4.30	6.12	**2.78**	4.12	4.65	3.36
		$M(P_{25}, P_{75})$	2.83(1.62, 4.76)	2.51(1.52, 5.08)	3.99(1.91, 6.37)	**2.36(1.40, 4.12)**	3.37(2.02, 6.39)	3.23(2.05, 5.37)	2.96(2.13, 5.88)	2.74(1.92, 4.60)
	HD (mm)	AVG ± SD	11.79 ± 6.02	10.13 ± 6.26	13.84 ± 8.34	**10.06 ± 6.13**	13.24 ± 8.18	15.26 ± 8.36	13.33 ± 8.13	12.25 ± 8.89
	DC (%)	AVG ± SD	**72.45 ± 19.18**	70.68 ± 18.92	69.11 ± 20.75	71.97 ± 18.02	70.38 ± 18.45	63.51 ± 18.47	69.50 ± 17.62	70.33 ± 19.28
CTV	MSD (mm)	SKEW	2.76	3.82	3.00	3.74	**2.68**	3.40	2.92	2.80
		$M(P_{25}, P_{75})$	**5.26(2.18, 8.82)**	6.24(4.12, 9.28)	5.48(2.96, 8.32)	5.46(3.08, 8.05)	7.36(5.12, 11.17)	7.06(3.86, 10.64)	6.38(4.05, 9.27)	5.74(3.20, 8.92)
	HD (mm)	AVG ± SD	**16.22 ± 10.01**	21.02 ± 9.15	17.80 ± 9.89	18.12 ± 9.16	21.95 ± 8.03	20.92 ± 9.30	17.77 ± 11.18	17.43 ± 12.94
	DC (%)	AVG ± SD	**71.06 ± 17.72**	60.46 ± 16.91	70.23 ± 16.59	70.61 ± 14.35	62.94 ± 16.15	60.89 ± 18.25	67.78 ± 18.68	68.93 ± 19.69

SKEW represented skewness coefficient, *M* represented the median, P_{25} represented the lower quartile, P_{75} represented the upper quartile, AVG represented the arithmetic mean, and SD represented the standard deviation

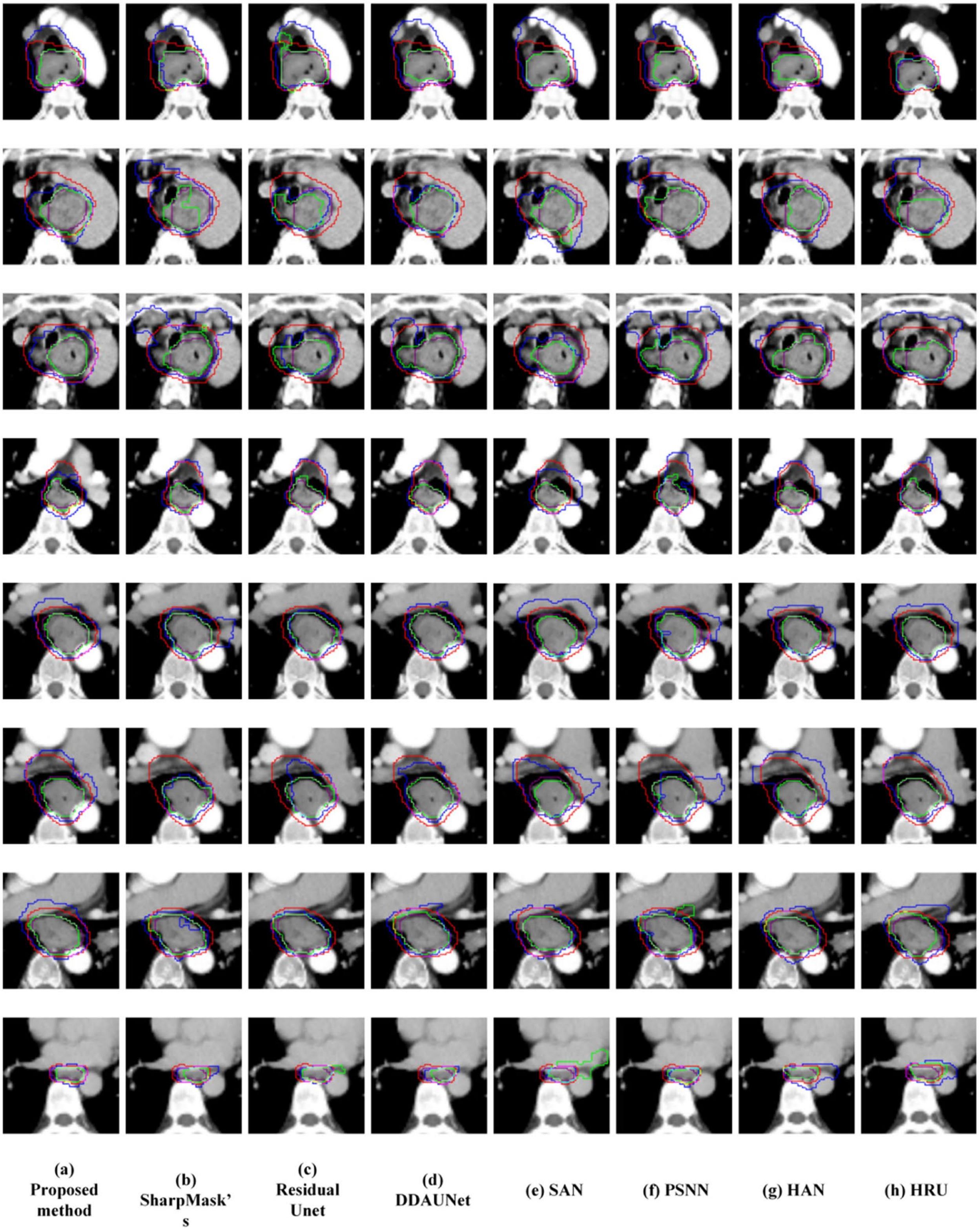
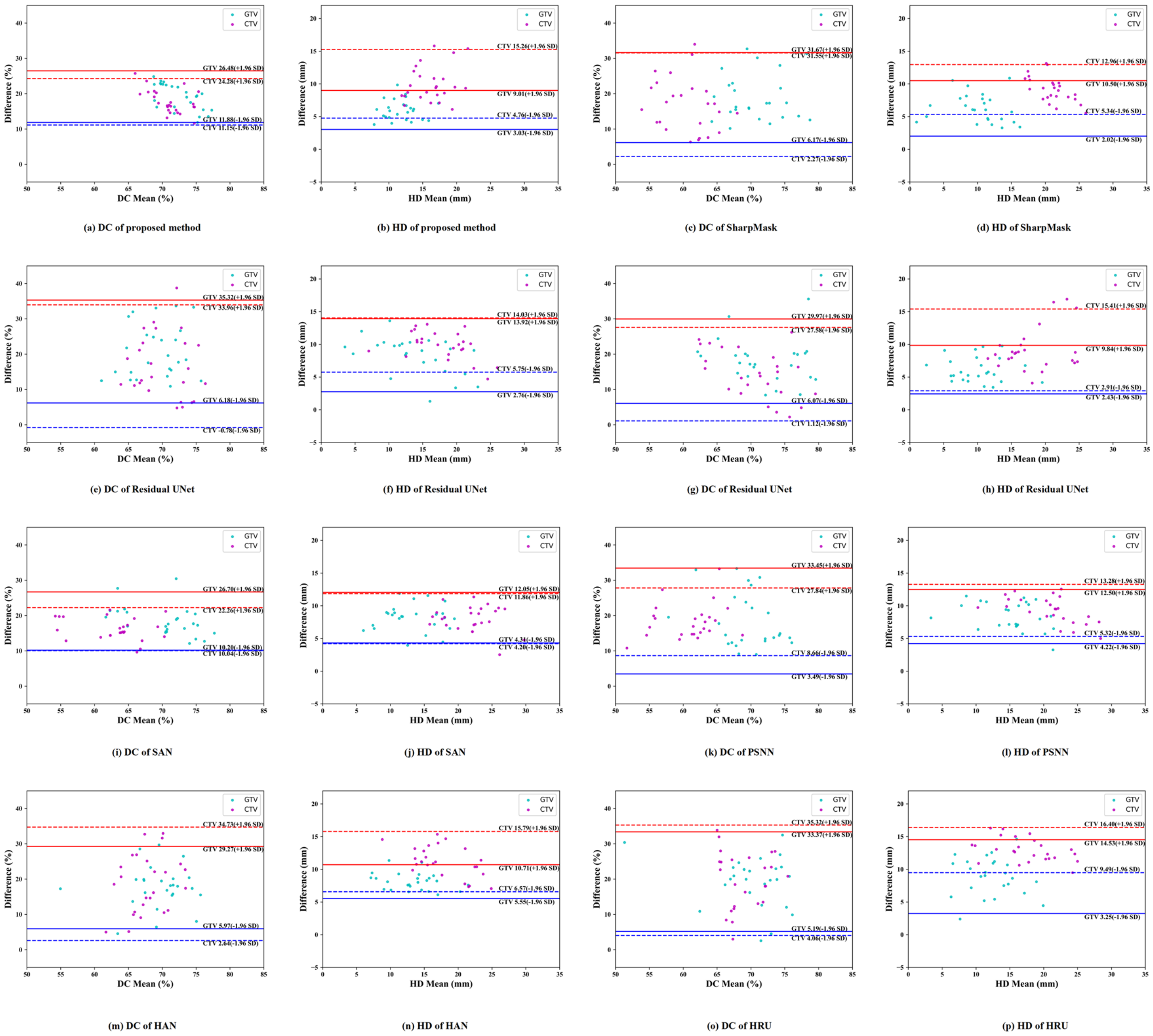


Fig. 7 Predictions example of the networks. Green and purple contours represented the prediction and ground truth of GTV, respectively; blue and red contours represented the prediction and ground truth of CTV, respectively; compound colors represented the overlap. Predictions are all transformed from regions to contours



(a) DC of proposed method (b) HD of proposed method (c) DC of SharpMask (d) HD of SharpMask

(e) DC of Residual UNet (f) HD of Residual UNet (g) DC of Residual UNet (h) HD of Residual UNet

(i) DC of SAN (j) HD of SAN (k) DC of PSNN (l) HD of PSNN

(m) DC of HAN (n) HD of HAN (o) DC of HRU (p) HD of HRU

Fig. 8 Bland–Altman analysis of the neural networks. Red and blue transverse lines represented the upper and lower bound of LoA, respectively. Soid line and dotted line represented GTV and CTV data set, respectively

In order to further reflect the clinical application value of the proposed method, the esophageal CT images shared by Lucchesi *et. al.* from Barretos Cancer Hospital in the TCIA database were quoted and predicted using the model trained in this study [55]. After filtration, a total of 36 cases of esophageal CT scans were available for prediction. Since the DFOV and spatial resolution of these cases were different, they were converted into common image format and uniformly adjusted to the resolution of 512×512 pixels, and then predicted using the trained model. Some examples of the prediction were shown in Fig. 9. The green and red contours represented the ground truth and prediction of GTV, and the blue and yellow contours represented the ground truth and prediction of CTV, respectively. The MSD, HD and DC in GTV were 3.98 ± 3.01 mm, 12.52 ± 9.65 mm, and 71.06 ± 20.43%, while the values of that in CTV were 5.45 ± 4.27 mm, 19.67 ± 12.14 mm, and 69.02 ± 23.83%.

In clinical practice, the application of the proposed network might be restricted by the performance of hardware and software, although the convolution layers has been decomposed to a large extent. Under the circumstances, the number of ECBs could be reduced to a reasonable degree and the fundamental quantity of the channels could be modified to an adaptive order. The adjusted network needs to be re-trained as a “tiny” version.

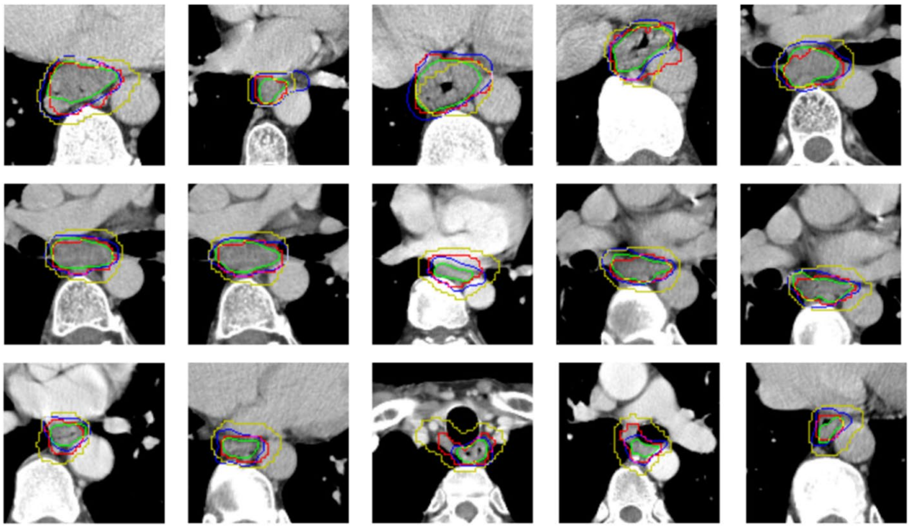


Fig. 9 Prediction examples from TCIA database. Green and red contours represented the ground truth and prediction of the esophageal GTV, respectively; blue and yellow contours represented the ground truth and prediction of the esophageal CTV, respectively

Conclusion

This paper proposed an esophageal segmentation method based on neural network ECAU in simulated CT images. This method structured a PCE module at the beginning of the network, which employed PCA method to expand the channels and suppress non-primary information. In ECAU, a series of ECBs was designed to replace the skip concatenation of original UNet, in order to better fuse the features from different encoder stages. In ECB, a WRA unit with large kernel convolution was assigned to enpand the receptive field, and the kernel was decomposed to reduce the amount of calculation. An SLA unit was assigned in parallel with WRA to supplement the capture capability of local attention. The FGRNL module was designed in IBN unit to fuse the extracted features after WRA and SLA, reconstructing the spectral response of the feature maps and strengthen the significant frequent component. After that, ablation experiment was carried out to reflect the performance of each module. Then, the prediction of the proposed network was compared with the published esophageal segmentation methods based on the same data set. The results showed a competitive performance of the proposed method in the segmentation task of esophageal GTV and CTV in simulated CT images. Nevertheless, the proposed method did not work excellently in the slices where the contrast between the esophagus and surrounding tissues was low. We planned to design a 3D segmentation networks which would have the potential capability to explore the extra features among the adjacent slices to enhance the performance of segmentation in the further research.

Acknowledgements
The authors are grateful to the Physical Technology Special Committee of Zheiiang Provincial Anti-Cancer Association.

Clinical trial number
Not applicable.

Authors' contributions
XL analyzed the data from the esophageal carcinoma patients, and was a major contributor in writing the manuscript. YZ collected the patients' CT images and delineated the target volumes. JZ and BL verified the delineation of the target volume. JY transferred the image format and interpreted the data. YZ and BL brought the idea of the study. HS offered the recommendations of the experiment. All authors read and approved the final manuscript.

Funding
This work was supported by the Doctoral Research Start-up Funds of Lishui People's Hospital (grant number 2024bs006), the Special Fund Project of the Physical Technology Special Committee of Zheiiang Provincial Anti-Cancer Association, and the Academic promotion program of Shandong First Medical University (2019LJ004).

Data availability
The data that support the findings of this study are available from the Ethics Committee of Lishui People's Hospital but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Ethics Committee of Lishui People's Hospital.

Declarations

Ethics approval and consent to participate
Approval was granted by the Ethics Committee of Lishui People's Hospital. Informed consent was obtained from all individual participants included in the study. All of the patients achieved normal treatment, but not for clinical trials or scientific research.

Consent for publication
Informed consent was obtained from all individual participants included in the study.

Competing interests

The authors declare no competing interests.

Received: 6 September 2024 Accepted: 26 November 2024
Published online: 18 December 2024

References

1. Jaffray DA. Image-guided radiotherapy: from current concept to future perspectives. Nat Rev Clin Oncol. 2012;9(12):688–99. https://doi.org/10.1038/nrclinonc.2012.194.
2. Dawson LA, Sharpe MB. Image-guided radiotherapy: rationale, benefits, and limitations. Lancet Oncol. 2006;7(10):848–58. https://doi.org/10.1016/S1470-2045(06)70904-4.
3. Verellen D, Ridder MD, Linthout N, et al. Innovations in image-guided radiotherapy. Nat Rev Cancer. 2007;7(12):949–60. https://doi.org/10.1038/nrc2288.
4. Hatch GF III, Wertheimer-Hatch L, Hatch KF, et al. Tumors of the esophagus. Nat Rev Cancer. 2000;24(4):401–11. https://doi.org/10.1007/s002689910065.
5. Daniel EE. Lower esophagus: structure and function. In: Sphincters: Normal Function-Changes in Disease. 1992. p. 49–66.
6. Hashizume M, Kitano S, Sugimachi K, et al. Three-dimensional view of the vascular structure of the lower esophagus in clinical portal hypertension. Hepatology. 1988;8(6):1482–7. https://doi.org/10.1002/hep.1840080603.
7. Stemkens B, Tijssen RH, de Senneville BD, et al. Image-driven, model-based 3D abdominal motion estimation for MR-guided radiotherapy. Phys Med Biol. 2016;61(14):5335–55. https://doi.org/10.1088/0031-9155/61/14/5335.
8. Lagendijk JJW, Raaymakers BW, Van den Berg CAT, et al. MR guidance in radiotherapy. Phys Med Biol. 2014;59(21):R349–69. https://doi.org/10.1088/0031-9155/59/21/R349.
9. Dietz B, Yip E, Yun JH, et al. Real-time dynamic MR image reconstruction using compressed sensing and principal component analysis (CS-PCA): demonstration in lung tumor tracking. Med Phys. 2017;44(8):3978–89. https://doi.org/10.1002/mp.12354.
10. Bjerre T, Crijns S, Af Rosenschold PM, et al. Three-dimensional MRI-linac intra-fraction guidance using multiple orthogonal cine-MRI planes. Phys Med Biol. 2013;58(14):4943–50. https://doi.org/10.1088/0031-9155/58/14/4943.
11. Rousson M, Bai Y, Xu C. Probabilistic minimal path for automated esophagus segmentation. In: Medical Imaging 2006: Image Processing. SPIE, 2006, 6144: 1361–1369. https://doi.org/10.1117/12.653657.
12. Huang T C, Zhang G, Guerrero T. Semi-automated CT segmentation using optic flow and Fourier interpolation techniques. Comput Meth Prog Bio. 2006, 84(2–3): 124–134. https://doi.org/10.1016%2Fj.cmpb.2006.09.003.
13. Fieselmann A, Lautenschläger S, Deinzer F. Esophagus segmentation by spatially-constrained shape interpolation. In: Bildverarbeitung für die Medizin 2008. Springer, 2008: 247–251. https://doi.org/10.1007/978-3-540-78640-5_50.
14. Feulner J, Zhou S K, Cavallaro A, et al. Fast automatic segmentation of the esophagus from 3D CT data using a probabilistic model. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009: 12th International Conference. 2009: 255–262. https://doi.org/10.1007/978-3-642-04268-3_32.
15. Kurugol S, Bas E, Erdogmus D, Centerline extraction with principal curve tracing to improve 3D level set esophagus segmentation in CT images. In, et al. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE. 2011;2011:3403–6. https://doi.org/10.1109/IEMBS.2011.6090921.
16. Yang J, Haas B, Fang R, et al. Atlas ranking and selection for automatic segmentation of the esophagus from CT scans. Phys Med Biol. 2017;62(23):9140. https://doi.org/10.1088/1361-6560/aa94ba.
17. You C, Yang J, Chapiro J, et al. Unsupervised wasserstein distance guided domain adaptation for 3d multi-domain liver segmentation. Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020. 2020: 155–163. https://doi.org/10.1007/978-3-030-61166-8_17.
18. You C, Zhao R, Liu F, et al. Class-aware adversarial transformers for medical image segmentation. Adv Neural Inf Process Syst. 2022, 35: 29582–29596. https://dl.acm.org/doi/abs/https://doi.org/10.5555/3600270.3602415.
19. You C, Xiang J, Su K, et al. Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation. International Workshop on Distributed, Collaborative, and Federated Learning. 2022: 3–16. https://doi.org/10.1007/978-3-031-18523-6_1.
20. You C, Zhao R, Staib L H, et al. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022: 639–652. https://doi.org/10.1109/TMI.2022.3161829.
21. You C, Zhao R, Staib L H, et al. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022: 639–652. https://doi.org/10.1007/978-3-031-16440-8_61.
22. You C, Dai W, Min Y, et al. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. International conference on information processing in medical imaging. 2023: 641–653. https://doi.org/10.1007/978-3-031-34048-2_49.
23. You C, Dai W, Min Y, et al. Action++: Improving semi-supervised medical image segmentation with adaptive anatomical contrast. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2023: 194–205. https://doi.org/10.1007/978-3-031-43901-8_19.
24. You C, Dai W, Min Y, et al. Implicit anatomical rendering for medical image segmentation with stochastic experts. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2023: 561–571. https://doi.org/10.1007/978-3-031-43898-1_54.
25. You C, Dai W, Liu F, et al. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. IEEE T Pattern Anal. 2024: 1–16. https://doi.org/10.1109/TPAMI.2024.3461321.
26. You C, Dai W, Min Y, et al. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. Adv Neural Inf Process Syst. 2024: 36. https://dl.acm.org/doi/abs/10.5555/3666122.3666559.
27. Zhu F, Gao Z, Zhao C, et al. Semantic segmentation using deep learning to extract total extraocular muscles and optic nerve from orbital computed tomography images. Optik. 2021;244: 167551. https://doi.org/10.1016/j.ijleo.2021.167551.
28. Liu X, Zhang D, Yao J, et al. Transformer and convolutional based dual branch network for retinal vessel segmentation in OCTA images. Biomed Signal Proces. 2023;83: 104604. https://doi.org/10.1016/j.bspc.2023.104604.
29. Liu X, Liu Q, Zhang Y, et al. TSSK-Net: Weakly supervised biomarker localization and segmentation with image-level annotation in retinal OCT images. Comput Biol Med. 2023;153: 106467. https://doi.org/10.1016/j.compbiomed.2022.106467.
30. Yang G, Geng P, Ma H, et al. Dwta-unet: Concrete crack segmentation based on discrete wavelet transform and unet. Proceedings of 2021 Chinese Intelligent Automation Conference. 2022: 702–710. https://doi.org/10.1007/978-981-16-6372-7_75.
31. Geng P, Tan Z, Wang Y, et al. STCNet: Alternating CNN and improved transformer network for COVID-19 CT image segmentation. Biomed Signal Proces. 2024;93: 106205. https://doi.org/10.1016/j.bspc.2024.106205.
32. Hao Z, Liu J, Liu J. Esophagus tumor segmentation using fully convolutional neural network and graph cut. In: Chinese Intelligent Systems Conference. Singapore. Springer, 2017: 413–420. https://doi.org/10.1007/978-981-10-6499-9_39.
33. Trullo R, Petitjean C, Nie D, et al. Fully automated esophagus segmentation with a hierarchical deep learning approach. In: 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, 2017: 503–506. https://doi.org/10.1109/ICSIPA.2017.8120664.
34. Chen S, Yang H, Fu J, et al. U-Net Plus: deep semantic segmentation for esophagus and esophageal cancer in computed tomography images. IEEE Access. 2019;7:82867–77. https://doi.org/10.1109/ACCESS.2019.2923760.
35. Diniz JOB, Ferreira JL, Diniz PHB, et al. Esophagus segmentation from planning CT images using an atlas-based deep learning approach. Comput Meth Prog Bio. 2020;197: 105685. https://doi.org/10.1016/j.cmpb.2020.105685.

36. Yousefi S, Sokooti H, Elmahdy MS, et al. Esophageal tumor segmentation in CT Images using a dilated dense attention Unet (DDAUnet). IEEE Access. 2021;9:99235–48. https://doi.org/10.1109/ACCESS.2021.3096270.
37. Tran M, Kim S, Yang H, et al. Esophagus segmentation in CT images via spatial attention network and STAPLE algorithm. Sensors. 2021;21(13):4556. https://doi.org/10.3390/s21134556.
38. Alam SR, Zhang P, Zhang SY, et al. Early Prediction of Acute Esophagitis for Adaptive Radiation Therapy. Int J Radiat Oncol. 2021;110(3):883–92. https://doi.org/10.1016/j.ijrobp.2021.01.007.
39. Jin DK, Guo DZ, Ho TY, et al. DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Med Image Anal. 2021;68: 101909. https://doi.org/10.1016/j.media.2020.101909.
40. Li D, Cheng Y, Guo Y, et al. Esophageal tissue segmentation on OCT images with hybrid attention network. Multimed Tools Appl. 2024;83(14):42609–28. https://doi.org/10.1007/s11042-023-16550-z.
41. Jian M, Tao C, Wu R, et al. HRU-Net: A high-resolution convolutional neural network for esophageal cancer radiotherapy target segmentation. Comput Meth Prog Bio. 2024;250: 108177. https://doi.org/10.1016/j.cmpb.2024.108177.
42. Karamizadeh S, Abdullah SM, Manaf AA, et al. An overview of principal component analysis. J Signal and Inform Process. 2013;04(03):173–5. https://doi.org/10.4236/jsip.2013.43B031.
43. Yao Y. A LUNet based on large kernel attention mechanism for image denoising. In: International Conference on Electronic Information Technology (EIT 2022). SPIE, 2022. https://doi.org/10.1117/12.2638621.
44. Lau K W, Po L M, Rehman Y A U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention design in CNN. Expert Syst Appl. 2024, 236: 121352.1–121352.15. https://doi.org/10.1016/j.eswa.2023.121352.
45. Ding X, Zhang X, Han J. Scaling up your kernels to 31x31: revisiting large kernel design in CNNs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 11963–11975. https://doi.org/10.1109/CVPR52688.2022.01166.
46. Ma N, Zhang X, Zheng H. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV). CVF, 2018: 116–131. https://doi.org/10.1007/978-3-030-01264-9_8.
47. Szegedy C, Vanhoucke V, Ioffe S. Rethinking the Inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
48. Woo S, Debnath S, Hu R. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2023: 16133–16142. https://doi.org/10.1109/CVPR52729.2023.01548.
49. Campbell FW, Robson JG. Application of Fourier analysis to the visibility of gratings. J Physiol. 1968;197(3):551. https://doi.org/10.1113/jphysiol.1968.sp008574.
50. Park J, Woo S, Lee J. BAM: bottleneck attention module. arXiv preprint arXiv:1807.06514. 2018. https://doi.org/10.48550/arXiv.1807.06514.
51. Wang Q, Wu B, Zhu P. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 11534–11542. https://doi.org/10.1109/CVPR42600.2020.01155.
52. Yang Z, Zhu L, Wu Y. Gated channel transformation for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 11794–11803. https://doi.org/10.1109/CVPR42600.2020.01181.
53. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 7132–7141. https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2913372.
54. Woo S, Park J, Lee J. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). CVF, 2018: 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.
55. Lucchesi FR, Aredes ND. The Cancer Genome Atlas Esophageal Carcinoma Collection (TCGA-ESCA) (Version 3). The Cancer Imaging Archive. 2016. https://doi.org/10.7937/K9/TCIA.2016.VPTNRGFY.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.