CAENet: Contrast adaptively enhanced network for medical image segmentation based on a differentiable pooling function

Computers in Biology and Medicine 167 (2023) 107578

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

CAENet: Contrast adaptively enhanced network for medical image segmentation based on a differentiable pooling function

Shengke Li [a,e], Yue Feng [a,*], Hong Xu [a,b], Yuan Miao [b], Zhuosheng Lin [a], Huilin Liu [c], Ying Xu [d], Fufeng Li [d,**]

[a] Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen, 529020, Guangdong, China
[b] Victoria University, Melbourne, 8001, Australia
[c] Basic Medical College, Shanghai University of Traditional Chinese Medicine, Shanghai, 201203, China
[d] Laboratory of TCM Four Processing, Shanghai University of TCM, Shanghai, 201203, China
[e] School of Engineering, Guangzhou College of Technology and Business, Foshan, 528100, Guangdong, China

ARTICLE INFO

Keywords:
Medical image
Semantic segmentation
Differentiable pooling function
Channel attention
Deep supervision

ABSTRACT

Pixel differences between classes with low contrast in medical image semantic segmentation tasks often lead to confusion in category classification, posing a typical challenge for recognition of small targets. To address this challenge, we propose a Contrastive Adaptive Augmented Semantic Segmentation Network with a differentiable pooling function. Firstly, an Adaptive Contrast Augmentation module is constructed to automatically extract local high-frequency information, thereby enhancing image details and accentuating the differences between classes. Subsequently, the Frequency-Efficient Channel Attention mechanism is designed to select useful features in the encoding phase, where multifrequency information is employed to extract channel features. One-dimensional convolutional cross-channel interactions are adopted to reduce model complexity. Finally, a differentiable approximation of max pooling is introduced in order to replace standard max pooling, strengthening the connectivity between neurons and reducing information loss caused by downsampling. We evaluated the effectiveness of our proposed method through several ablation experiments and comparison experiments under homogeneous conditions. The experimental results demonstrate that our method competes favorably with other state-of-the-art networks on five medical image datasets, including four public medical image datasets and one clinical image dataset. It can be effectively applied to medical image segmentation.

1. Introduction

MEDICAL image segmentation has various applications and research values for disease diagnosis and analysis [1–3]. Conventional methods based on thresholding and morphological manipulation have achieved certain application results in medical image segmentation. However, their limitations, such as dependence on prior knowledge and strict application conditions, lead to unsatisfactory generalization performance. With the successful application of deep learning methods in computer vision, the semantic segmentation method based on the Convolutional Neural Network (CNN) has been widely used in the automatic segmentation of medical images owing to its low dependence on prior knowledge and powerful ability of feature learning. These methods excel in learning and extracting advanced semantic information from images, improving the distinction of different target regions. In addition, in comparison to traditional methods, these methods have better robustness and adaptability and can handle various image segmentation tasks under complex scenarios. Therefore, CNN-based semantic segmentation technology has become a vital tool in medical image analysis.

The CNN includes multiple layers, including but not limited to the convolutional layer, nonlinearity layer, pooling layer, and fully connected layer [4]. The Fully Convolutional Network (FCN) [5] revolutionized the fully connected layers of classification networks with convolutional layers, creating an encoder-decoder structure for fully convolutional semantic segmentation. This enables end-to-end, pixel-wise segmentation. Subsequently, FCNs have been widely applied in various segmentation tasks such as brain tumor segmentation, skin

* Corresponding author.
** Corresponding author.
E-mail addresses: J002443@wyu.edu.cn (Y. Feng), li_fufeng@aliyun.com (F. Li).

https://doi.org/10.1016/j.compbiomed.2023.107578
Received 18 March 2023; Received in revised form 3 October 2023; Accepted 15 October 2023
Available online 17 October 2023

lesion segmentation, etc. [6]. However, FCN only utilizes high-level semantic features, which makes it challenging to precisely localize objects, resulting in relatively coarse segmentation results and the loss of fine details. Unlike general image segmentation tasks, medical images typically contain noise and exhibit blurred boundaries, requiring the model not only to detect and recognize high-level semantic features but also to rely on low-level features for accurate boundary delineation. Based on FCN, U-Net [7] introduces concatenation structures between the corresponding encoder and decoder layers. In U-Net, skip connections are utilized to merge the feature maps from the encoding end with those at the decoding end, effectively combining low-resolution and high-resolution image features. This design enables end-to-end segmentation performance without any pretraining. Its simplicity and efficiency render it a popular choice for medical applications. Consequently, U-Net has become the benchmark for medical image segmentation. To date, it has been widely used in various biomedical image segmentation tasks, such as cardiac segmentation by magnetic resonance, optic disc and cup segmentation in retinal fundus images, and organ segmentation in Computed Tomography (CT) images. Numerous advanced medical image segmentation networks have been improved based on this architecture, including SegNet [8], U-Net++ [9], AttU-Net [10], CPFNet [11], and CA-Net [12], among others. These improvements attest to the effectiveness of the symmetric coding structure. However, these methods directly input the input image into the network of symmetrical coding and decoding structure for calculation, while our method sets a trainable contrast enhancement layer in the first layer of the network, and then performs segmentation on the enhanced image through the network of symmetrical coding and decoding structure.

The semantic information in medical images may vary depending on the organ or tissue being examined, making it challenging to accurately capture this information during segmentation. As a result, many researchers have developed advanced network structures to address these challenges and improve the accuracy of medical image segmentation. U-Net++ was designed to address the large semantic gap between the features of U-Net's skip connections by replacing them with nested and dense skip connections. The improvement was validated on colon polyp, liver, cell nuclei, and lung nodule datasets. CPFNet proposed a global pyramid guidance module that provides different levels of global context information for the decoder to dynamically fuse multiscale context information in high-level features. Experimental results show that CPFNet is highly competitive compared with other state-of-the-art methods on four different challenging tasks. CE-Net [13] incorporates a dense atrous convolution block that captures deeper and wider context features by fusing cascaded paths and a residual multikernel pooling block encoding global context information at multi-scale receptive fields. However, medical images can often contain noise, artifacts, and other irrelevant information. Direct feature reuse between convolutional layers might lead to negative transfer of feature knowledge.

As a result, attention mechanisms have been increasingly used in medical image segmentation to help focus on the most relevant information in the feature map, enabling more accurate segmentation of varying sizes and shapes of target structures. AttU-Net designed an attention gate block to suppress the unrelated information in the feature map, improving the accuracy of multiclass image segmentation on two large CT abdominal datasets. However, a single type of attention may ignore other important information. To address this issue, CA-Net constructed a comprehensive attentional network that combines spatial, channel, and scale attention to focus more on the foreground area, resulting in improved performance compared with the U-Net on skin lesion dataset from ISIC 2018. DANet [14] appended two types of attention modules on top of dilated FCN to model semantic interdependencies in spatial and channel dimensions, achieving state-of-the-art results on three scene segmentation datasets, and was later being applied to medical image segmentation tasks [15]. MEA-Net [16] focused on edge information extraction and introduced a new multilayer edge attention module to address the functional defects of the current attention mechanism system. The module performs spatial compression through global average pooling and further integrates features from shallow encoding layers. It achieved the best segmentation results on three public medical image datasets, including retinal vessels, lung and tongue images, as well as a clinical tongue dataset. However, complex attention structures and multi-level utilization of attention can easily lead to redundant use of computing resources and model parameters. Our method considers eliminating redundant features through a lightweight channel attention during encoding, and utilizes spatial attention to focus on details such as position and shape when the first-layer encoded features are passed to the symmetrical decoding layer.

In CNN-based semantic segmentation, the process involves two stages: feature extraction and resolution restoration. Pooling is employed to reduce resolution, effectively ignoring noise and decreasing the number of parameters to speed up training, leading to improved target recognition. However, the lost spatial information during pooling is not fully recovered through upsampling. To address this, aside from the previously mentioned method of cross-layer feature reuse, researchers have designed dedicated information compensation modules [13] and replaced pooling layers with convolutions of stride 2 [17]. Furthermore, using improved pooling functions is also a direct and effective approach to exploring solutions for this issue. Ms RED [18] introduced a novel pooling module (Soft-pool) to medical image segmentation for the first time, retaining more helpful information when downsampling and getting better segmentation performance. Inspired by this, we replace max pooling in the vertical encoding and transmission process with its differentiable approximation function, and use a trainable parameter to control the output size of pooling at different layers.

In this study, we propose a Contrast-Adaptive Enhanced Network (CAENet) for semantic segmentation of medical images. The network is built upon a symmetric encoder-decoder structure, where an improved pooling function replaces the standard max pooling operation to enhance the continuity of semantic information transmission. The contrast enhancement module in the network actively enhances the quality of input images before encoding, aiming to improve the information quality of the images. An improved channel attention mechanism is employed to suppress redundant features during the vertical encoding propagation, and a spatial attention mechanism is used to focus on detailed information such as target positions. Additionally, the Deep Supervision strategy (DeepSup) is incorporated to further refine object delineation and expedite model convergence. The main contributions of this study are summarized as follows.

1) We propose an Adaptive Contrast Augmentation (ACA) module that can adaptively enhance the contrast of images during training.
2) We reconsidered the Max Pooling function (MaxPool) from the perspective of the differentiable approximation of the maximum and introduced the Smooth Maximum Pooling function (SMPool) to replace the maximum pooling function in the network, to reduce the information loss caused by downsampling in the coding process.
3) Inspired by FCA [19] and ECA modules, we designed the Frequency-Efficient Channel Attention (FECA) module to enhance the useful channel features in the encoding stage.
4) We extensively evaluated our proposed network, CAENet, on five distinct datasets, including four public medical image datasets and one clinical image dataset, and the evaluation demonstrated superior performance.

2. Related works

2.1. Attention mechanisms

Attention mechanisms have achieved great success in many visual

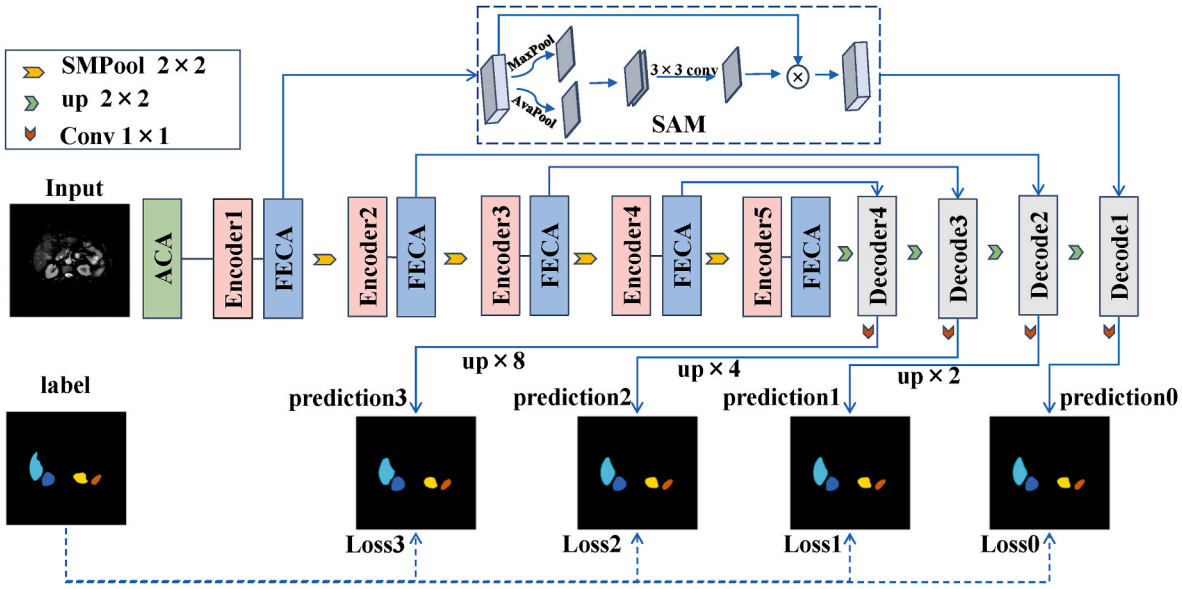

Fig. 1. Overview of the CAENet. Each encoding layer consists of an encoder and FECA module, and every encoder includes two 3 × 3 convolutions with a BN layer and ReLU activation function.

tasks, including image classification, object detection, semantic segmentation, video understanding, image generation, 3D vision, multimodal tasks, and self-supervised learning. Among these, SENet [20] stands out with its Squeeze and Excitation (SE) block, which recalibrates channel features. This approach has been widely adopted and has been used as a foundation for many advanced attention modules. FCANet [19] is based on the understanding that using average pooling on its own is not sufficient to represent all feature information when extracting channel information. SE is reconceptualized from the perspective of frequency, and Frequency Channel Attention (FCA) is proposed to enrich feature information by introducing more frequency components. ECA-Net [21] demonstrates that appropriate cross-channel interaction can significantly reduce the model complexity while maintaining performance. It introduces an Efficient Channel Attention (ECA) that replaces the fully connected layer in the SE module with a 1D convolution operation without dimensionality reduction, resulting in improved performance with fewer parameters. CBAM [22] incorporates both channel attention module and Spatial Attention Module (SAM) in serial to achieve higher accuracy and lower error rate. Coordinate attention [23] decomposes channel attention into two 1D feature coding processes that capture long-range dependence in one spatial direction and preserve precise location information in the other.

These attention mechanisms have shown promising results in various medical image segmentation tasks and have contributed to the development of more sophisticated and accurate models for this task [24].

2.2. Contrast enhancement methods

Medical images, constrained by the performance of imaging devices, may suffer from issues such as low contrast and image blurriness, which pose challenges for further processing. Hence, contrast enhancement of the original images becomes crucial. In most cases, the contrast of the raw medical images is exceedingly low, which hinders the extraction of accurate features using the original network [25]. Therefore, contrast enhancement methods can be considered to improve the segmentation performance of the network. Depending on the treatment range, the commonly used contrast enhancement methods can be divided into two main groups: global and local approaches.

Within the global enhancement methods, Linear Contrast Stretch (LCS) and Histogram Equalization (HE) are classical global image enhancement methods. While LCS linearly adjusts the dynamic range of the image, HE aims to uniformize the histogram distribution of the input image. However, both methods generally suffer from undersaturation and oversaturation, resulting in inferior-quality images [26].

On the other hand, local enhancement methods, such as Adaptive Histogram Equalization (AHE) and Contrast Limited Adaptive Histogram Equalization (CLAHE), are widely employed. AHE divides the image into multiple subregions and applies HE independently to each of them. CLAHE, an improved version of AHE, addresses issues related to abrupt transitions and over-enhancement. Another notable local approach is Adaptive Contrast Enhancement (ACE), which divides the image into high- and low-frequency components. It enhances the high-frequency part by calculating a gain coefficient, subsequently reconstructing the low-frequency component to obtain an enhanced image.

From the perspective of linear enhanced image contrast, we constructed a convolution module for enhancing the input image combined with the adaptability of convolution training.

2.3. Pooling methods

Pooling layers are crucial in deep learning tasks such as segmentation and classification. On one hand, it reduces the computational cost of the network and increases its efficiency. On the other hand, it can increase the receptive field of the convolution during encoding [27]. Max pooling and average pooling are common pooling functions in various networks, with max pooling being most commonly used in semantic segmentation networks [7,10–12,16]. However, max pooling makes parts other than the position corresponding to the maximum zero during backpropagation, resulting in the loss of crucial information [13,16].

To solve this challenge, retaining important information during downsampling becomes crucial. Owing to the combination of the networks discussed above, the information loss caused by the pooling process is mostly remedied by combining existing methods, including the design of specialized modules for information compensation [13] and replacing the pooling layer with convolution with a stride of 2 [17]. However, direct optimization of the pooling function to alleviate this loss has been relatively rare.

Taking into account the combination of max pooling and average pooling, researchers have explored innovative approaches. For instance, Yu et al. [28] applied both maximum and average pooling according to probability, and Lee et al. [29] designed a pooling based on a decision tree to combine the max pooling and average pooling function.

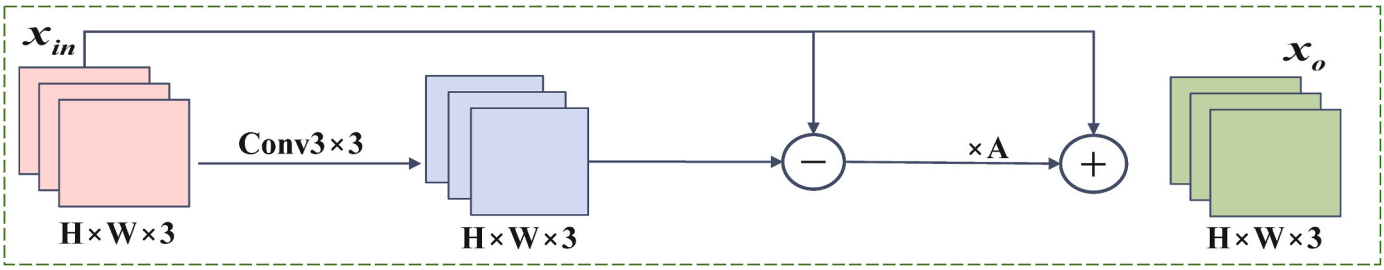

Fig. 2. Overall structure of the ACA module. *H* and *W* denote the height and width of the image, respectively. A is a constant that weights high-frequent features by a default value of 2.

Alexandros et al. [27] introduced the Soft Pooling function (SoftPool) by calculating the exponentially weighted sum of activations, which achieved performance improvement in the classification task. Ms RED [18] replaced the max pooling in the network with SoftPool in the segmentation of skin lesion datasets, proving the effectiveness of SoftPool in the task of medical image segmentation.

In contrast to developing specialized modules for information compensation, we believe that a suitable pooling function is more convenient owing to fewer parameters and complexity. Therefore, we design a differentiable pooling function to replace max pooling in the encoding layer, which reduces the information loss of max pooling.

2.4. Deep supervision

In the field of medical image segmentation, the deep supervision strategy proves effective in aiding networks to capture multi-level features within images, making it a common training strategy. The recent study by MS-Dual-Guided [15] observed that introducing additional supervision at each scale enhanced the segmentation performance of the proposed model. This approach also facilitated deep supervision to allow the model to effectively capture the shapes and sizes of distinct object categories.

In the context of salient object segmentation tasks, U2-Net [30] employed deep supervision to capture the multi-level structure of images, enhancing its ability to focus on both low-frequency and high-frequency regions. PraNet [31] applied deep supervision to achieve segmentation from coarse to fine-grained segmentation strategies. GFANet [6] used deep supervision to progressively localize and refine objects for skin lesion image segmentation tasks. In these studies, the deep supervision approaches employed in the above-mentioned studies all involve directly upsampling the output maps from the last encoding layer and all decoding layers to match the size of the label. Subsequently, losses are calculated separately for each and combined as the total model loss.

However, some researchers argue that upsampling the output maps may lead to unnecessary information loss. An alternative approach has been proposed where the label is downsampled to match the size of each output map, and then the losses are computed individually and aggregated as the final loss, as demonstrated in Ref. [32].

3. Methods

3.1. Network architecture

The overall architecture of the CAENet is shown in Fig. 1, starting with the ACA module followed by the encoding and decoding layers. FECA is used to assign weights to different channel features from the output of the encoder to enhance useful features. To reduce the loss caused by max pooling, downsampling between encoding layers is replaced by SMPool, and SAM is subsequently introduced at the first skip connection between encoding and decoding to select useful spatial information. The upsampling method at the decoding end is the bilinear interpolation. The outputs of each decoder (prediction 0–3) are supervised during training, where each pair of the prediction and the ground truth label is used to compute the total loss. During inference, the prediction 0 is considered as the final output.

3.2. Adaptive Contrast Augmentation module (ACA)

When the contrast of medical images is inferior, the boundaries between the target and background and between targets are not clear, which increases the difficulty of feature extraction. Therefore, the high-frequency components in the image can be enhanced before feature extraction. The global contrast enhancement method enhances the entire image, which is suitable for cases where the gray values of useful objects are close; nonetheless, the interference and noise information are simultaneously enhanced. Whereas the local contrast enhancement method separately enhances the local features of the image and is suitable for images with large differences in gray values, it may over-enhance the gray peak region.

Based on the concept of linear enhancement of LCS and high-frequency extraction of ACE, we propose the ACA module, a contrastive enhancement module designed to adaptively enhance the high-frequency information within an image. As shown in Fig. 2, we first smooth the raw images x_{in} through a trainable convolutional filter, such that the filter parameters can be optimized as the network is trained. The smoothed image is subsequently subtracted by x_{in} to obtain the high-frequency component of the image. A controlled multiplicative factor, A, is thereafter used to control the multiplier of the enhancement. Finally, the augmented component is superimposed on x_{in} to obtain the augmented image x_o. In this process, adaptability is manifested in the use of image local features for processing and the update of convolution kernel parameters along with model training.

3.3. Smooth Maximum Pooling function (SMPool)

Max pooling reduces the amount of data by considering the maximum value, usually by selecting the largest pixel value from a subregion of the input feature map as the result of pooling. The MaxPool can be defined as follows:

$$\max(x_1, \cdots, x_N), \tag{1}$$

where $x_1, \cdots, x_N$ denote all values for a given pooled kernel region, and N is the number of the pooled pixels.

Since $\max(x_1, \cdots, x_N)$ is not differentiable, losing information during network training is easy, which is not conducive to identifying small objects and details. Therefore, its approximate differentiable function can be considered as a substitute. According to the Kreisselmeier–Steinhauser (KS) function [33],

$$\max(x_1, \cdots, x_N) = \lim_{\beta \to \infty} \frac{1}{\beta} \ln\left(\sum_{i=1}^{N} e^{\beta x_i}\right). \tag{2}$$

Accordingly, by L'Hopital's rule of the limit,

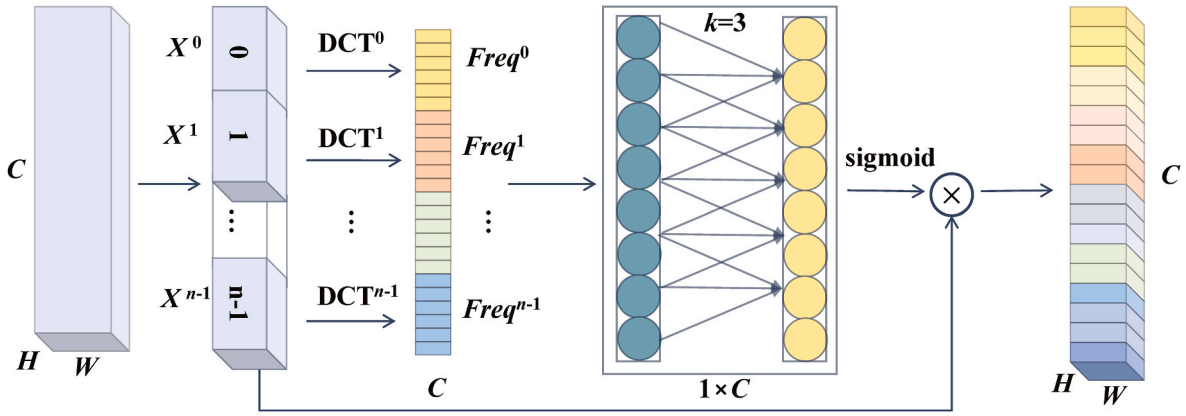

Fig. 3. Illustration of the FECA module. K denotes the size of the convolution kernel.

$$\lim_{\beta\to\infty}\frac{1}{\beta}\ln\left(\sum_{i=1}^{N}e^{\beta x_i}\right)=\lim_{\beta\to\infty}\frac{\sum_{j=1}^{N}x_j e^{\beta x_j}}{\sum_{i=1}^{N}e^{\beta x_i}}. \tag{3}$$

We can obtain two differentiable approximation functions SM1 and SMPool2 for the maxima:

$$\mathrm{SM1}=\frac{1}{\beta}\ln\left(\sum_{i=1}^{N}e^{\beta x_i}\right), \tag{4}$$

$$\mathrm{SMPool2}=\frac{\sum_{j=1}^{N}x_j e^{\beta x_j}}{\sum_{i=1}^{N}e^{\beta x_i}}. \tag{5}$$

According to (5), $x_{\min}\le \mathrm{SMPool2}\le x_{\max}$ can be inferred, and the equality sign holds if and only if all x's are equal.

Since β is in the denominator, when $\beta\to 0$,

$$\lim_{\beta\to 0}\mathrm{SM1}=\lim_{\beta\to 0}\frac{1}{\beta}\ln\left(\sum_{i=1}^{N}e^{\beta x_i}\right)=\infty. \tag{6}$$

In this case, SM1 is divergent, which is not conducive to the correct expression of features when used for pooling. When $\beta>0$,

$$x_{\max}\le \mathrm{SM1}\le\frac{\ln N}{\beta}+x_{\max}. \tag{7}$$

At this point, the value of SM1 is larger than the maximum value of the pooling region. Let $\mathrm{SM1}-\frac{\ln N}{\beta}$ to construct a new form SMPool1:

$$\begin{aligned}\mathrm{SMPool1}&=\mathrm{SM1}-\frac{\ln N}{\beta}\\&=\frac{1}{\beta}\ln\left(\sum_{i=1}^{N}e^{\beta x_i}\right)-\frac{\ln N}{\beta}\\&=\frac{1}{\beta}\ln\left(\frac{1}{N}\sum_{i=1}^{N}e^{\beta x_i}\right),\qquad \beta>0.\end{aligned} \tag{8}$$

According to (8), it is true that $x_{\min}\le\mathrm{SMPool1}\le x_{\max}$, $\lim_{\beta\to 0}\mathrm{SMPool2}=\lim_{\beta\to 0}\mathrm{SMPool2}=x_{mean}$, and $\lim_{\beta\to\infty}\mathrm{SMPool1}=\lim_{\beta\to\infty}\mathrm{SM1}=\lim_{\beta\to\infty}\mathrm{SMPool2}=x_{\max}$.

Accordingly, we can obtain two differentiable approximate functions of the MaxPool: SMPool1 and SMPool2. According to (5) and (8), compared with that of SMPool2, the calculation of SMPool1 is simple, and both are equally differentiable of higher order. When $\beta=0$, SMPool2 = AvgPool, and when $\beta=1$, SMPool2 = SoftPool.

There is only one parameter trained to automatically control the bias of SMPool1 or SMPool2, and due to $\beta\to\infty$, SMPool1 is $x_{\max}$, $\beta>0$ is set when it is used in the model. In our model, for each feature map, the value of β is determined, larger activation values will exert a stronger influence on the output.

3.4. Frequency-Efficient Channel Attention (FECA)

Attention mechanisms have proven to be valuable for enhancing the performance of deep CNNs. Building upon this foundation, our aim is to develop an effective channel attention mechanism that doesn't introduce excessive parameters, which could lead to model overfitting. Thus, we introduce the FECA, as illustrated in Fig. 3. FECA facilitates channel communication by employing a one-dimensional convolution with a kernel size of 3, all while avoiding the reduction of channel dimension. However, in situations with limited channel information, channel interactions may be insufficient. Drawing inspiration from the approach used in FCA, we utilize a 2D Discrete Cosine Transform (2DDCT) to compute multi-frequency component information for the channels. This incorporating multiple frequency aspects allows us to extract compressed channel details more effectively.

First, the input image I is divided into n equal parts denoted as $I^0, I^1, \cdots, I^{n-1}$ according to the channel, in which $I^i\in \mathrm{R}^{C/n\times H\times W}$. On each channel component, the frequency components $Freq^0, Freq^1, \cdots, Freq^{n-1}$ are calculated according to (9).

$$Freq^i=\mathrm{2DDCT}^{u_i,v_i}\left(I^i\right)=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}I^i_{:,h,w}H^{u_i v_i}_{h,w}, \tag{9}$$

$$H^{u_i,v_i}_{h,w}=\cos\left(\frac{\pi h}{H}\left(u_i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(v_i+\frac{1}{2}\right)\right), \tag{10}$$

$$\begin{aligned}&s.t.\ i\in\{0,1,\cdots,n-1\},\\&\quad h\in\{0,1,\cdots,H-1\},\\&\quad w\in\{0,1,\cdots,W-1\},\end{aligned}$$

where 2DDCT denotes the 2D discrete cosine transform, H and W denote the height and width of the image, respectively, u and v represent component subscript of 2DDCT as the same as [19] by default, and n is set to 16.

Subsequently, the obtained values of each frequency component are respliced according to the original channel division, and y is activated by one-dimensional convolution without dimensionality reduction to obtain the channel weight ω, as shown in (11), in which $\mathrm{C1D}_3$ represents the one-dimensional convolution operation with convolutional kernel size 3.

$$\omega=\mathrm{sigmoid}\left(\mathrm{C1D}_3\left(\mathrm{cat}\left(\left[Freq^0, Freq^1, \cdots, Freq^{n-1}\right]\right)\right)\right) \tag{11}$$

SE utilizes average pooling to extract channel information and employs a dimension-reducing followed by dimension-increasing fully connected layer for channel feature interaction. This addition

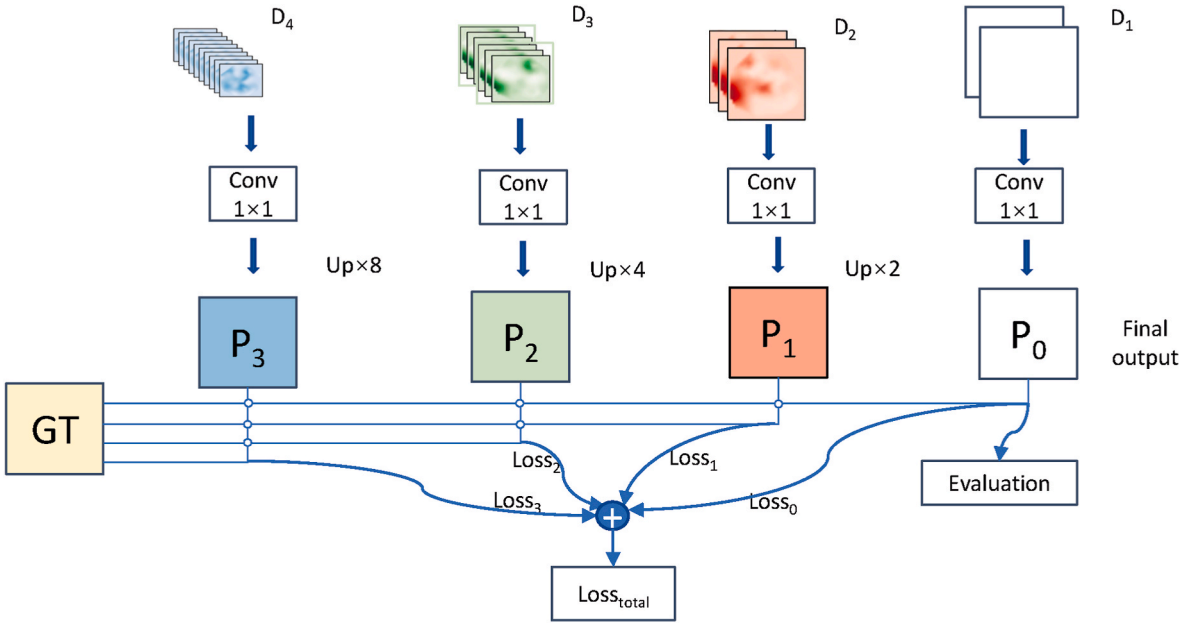

Fig. 4. Deep Supervision Method. Where D_i represents the output map of the *i*-th decoder, *i is in* {1, 2, 3, 4}; P_j denotes a one-channel feature map of the same size as GT for the *j*-th decoder, *j* is in {0, 1, 2, 3}; Up × *n* signifies bilinear upsampling by a factor of *n*, *n* is in {2, 4, 8}.

contributes an increment of 76.288K parameters to the proposed model. ECA, an improvement over SE, uses one-dimensional convolutions for effective local cross-channel interaction, reducing computational overhead and adding only 0.023K parameters to the proposed model. The kernel size of the one-dimensional convolution is determined by the number of input feature channels. Building upon SE, FCA regards the channel representation problem as a compression process using frequency analysis, which generalizes the existing channel attention mechanism in the frequency domain. Since 2DDCT calculations do not involve trainable parameters, the parameter quantity introduced by FCA remains the same as that of SE. Meanwhile, our proposed FECA combines the strengths of FCA and ECA. It calculates different frequency components for various channel components and achieves channel interaction through a non-reduced one-dimensional convolution with a kernel size of 3. Remarkably, this method incurs only 0.015K parameters in the model, making it a more lightweight and efficient attention module.

3.5. Deep supervision strategy

In medical image segmentation, deep supervision strategies play a crucial role in enhancing segmentation performance, offering adaptability to various types of medical image data and task requirements.

We assume that the bottleneck layer has a high number of channels, which necessitates a significant number of parameters for channel transformation. Furthermore, direct upsampling of the bottleneck layer, as demonstrated in Ref. [32], may lead to unnecessary loss of target information. In contrast to the conventional approach of supervising using the bottleneck layer (the last encoding layer in Fig. 1) and all the decoder's outputs [6,19,30–32], we supervise the output of each decoder individually, the deep supervision process is shown in Fig. 4.

First, the output maps of each decoder with different channel dimensions in Fig. 1 are unified to a single channel through a 1 × 1 convolution. Second, the output maps are resized to match the size of the Ground Truth (GT) through bilinear interpolation. Finally, during model training, losses are computed separately for each map P_i and the GT, and the aggregated loss serves as the final model loss. The ultimate prediction result is obtained from P_0, and it is combined with the GT to assess the segmentation performance of the model.

4. Experiments and analysis

4.1. Datasets

Five datasets, including CHAOS-T1, CHAOS-T2, Lung, Tongue, and Clinical-Face, were used to evaluate the performance of the different methods, as detailed below.

CHAOS-T1 and CHAOS-T2: The two abdominal MRI datasets are from the Combined Healthy Abdominal Organ Segmentation (CHAOS) Challenge [34–36]. We focus on the segmentation of abdominal organs (spleen, liver, and kidneys) on MRI T1DUAL in phase and T2SPIR. Each slice of both datasets has a resolution of 256 × 256 pixels. We chose the original training dataset for our experiments, randomly splitting its 2D slices into training (80 %), validation (10 %), and test set (10 %).

Lung: This dataset includes 267 images and their respective labels. These 2D CT lung images are acquired from the Lung Nodule Analysis (LUNA) competition. The size of each image is 512 × 512. During the experiments, images were randomly split into the training, validation, and test sets with a ratio of 6:2:2. The source data is available at https://www.kaggle.com/datasets/kmader/finding-lungs-in-ct-data.

Tongue: Tongue images were acquired from the TongeImageDataset. This dataset comprises 300 images with their respective labels published by BioHit. The size of each tongue image is 768 × 576 pixels. These images have been resized to 512 × 512 pixels and are randomly split into the training, validation, and test sets with a ratio of 6:2:2 during experiments. The source data is available at https://github.com/BioHit/TongeImageDataset.

Clinical-Face: This dataset is acquired from the Shanghai University of Traditional Chinese Medicine, Shanghai, China. Informed consent has been obtained for the release of the identifying images. Face images were captured by specialized equipment in an open environment, and five key-organ mapping regions (forehead-heart, left cheek-liver, nose-spleen, right cheek-lung, and jaw-kidney) were annotated by clinical experts. There are 180 images with a dimension of 1080 × 1440 in the original dataset; nonetheless, it is resized to 224 × 320 due to computational limitations. In our experiments, we use 80 % of the dataset for training, 10 % for validation, and 10 % for testing.

Table 1
Results of contrast experiments on CHAOS-T1 dataset (mean ± standard deviation). The best results are shown in bold. The backbone of FCN and DANet are ResNet101 and ResNet50, respectively.

Network	mIoU (%)	mF1 (%)	PA (%)	mRe (%)	HD (mm)
FCN [5]	73.63 ± 0.91	84.67 ± 0.61	98.99 ± 0.06	91.99 ± 0.54	1.45 ± 0.10
SegNet [8]	42.66 ± 29.30	92.77 ± 1.03	98.79 ± 0.47	47.64 ± 33.66	1.73 ± 0.33
U-Net [7]	86.04 ± 0.38	92.47 ± 0.22	99.49 ± 0.05	93.27 ± 0.89	1.24 ± 0.10
U-Net++ [9]	73.26 ± 3.45	84.33 ± 2.35	99.09 ± 0.11	87.81 ± 2.17	1.85 ± 0.14
AttU-Net [10]	86.73 ± 0.62	92.87 ± 0.36	99.49 ± 0.07	92.86 ± 1.19	1.23 ± 0.12
DANet [14]	79.34 ± 0.77	88.41 ± 0.48	99.24 ± 0.08	93.16 ± 0.53	1.37 ± 0.08
CPFNet [11]	82.76 ± 0.64	90.52 ± 0.38	99.37 ± 0.07	92.03 ± 0.50	1.30 ± 0.09
MEA-Net [16]	82.43 ± 0.98	90.30 ± 0.60	99.41 ± 0.06	88.33 ± 0.80	1.34 ± 0.06
CAENet (ours)	**88.21 ± 0.38**	**93.72 ± 0.21**	**99.55 ± 0.04**	**93.99 ± 0.14**	**1.15 ± 0.09**

Table 2
Results of contrast experiments on CHAOS-T2 dataset (mean ± standard deviation). The best results are shown in bold. The backbone of FCN and DANet are ResNet101 and ResNet50, respectively.

Network	mIoU (%)	mF1 (%)	PA (%)	mRe (%)	HD (mm)
FCN [5]	77.50 ± 0.65	87.30 ± 0.42	98.84 ± 0.03	92.80 ± 0.64	1.57 ± 0.02
SegNet [8]	57.27 ± 26.65	92.45 ± 0.26	98.83 ± 0.47	61.67 ± 28.81	1.76 ± 0.35
U-Net [7]	86.30 ± 1.11	92.63 ± 0.65	99.31 ± 0.03	91.95 ± 1.25	1.42 ± 0.01
U-Net++ [9]	74.82 ± 3.12	85.47 ± 2.06	98.85 ± 0.14	85.65 ± 1.84	2.01 ± 0.16
AttU-Net [10]	86.87 ± 1.11	92.96 ± 0.64	99.31 ± 0.06	92.17 ± 1.02	1.41 ± 0.05
DANet [14]	81.60 ± 1.00	89.85 ± 0.61	99.11 ± 0.06	92.18 ± 0.67	1.51 ± 0.02
CPFNet [11]	83.24 ± 1.02	90.84 ± 0.62	99.18 ± 0.05	91.10 ± 0.62	1.50 ± 0.05
MEA-Net [16]	84.18 ± 2.41	91.37 ± 1.45	99.19 ± 0.11	90.07 ± 2.61	1.56 ± 0.06
CAENet (ours)	**89.95 ± 0.86**	**94.70 ± 0.48**	**99.49 ± 0.04**	**94.33 ± 1.10**	**1.26 ± 0.03**

4.2. Experimental configuration

Experimental settings. All experiments were performed in the PyTorch 1.8.0 library with an NVIDIA RTX 6000 GPU. We train all the networks using the stochastic gradient descent optimizer with a mini-batch of size 16, and the momentum and weight decay are set to 0.9 and 5e-4, respectively. The learning rate policy adopts StepLR (step_size = 200, gamma = 0.5) with an initial learning rate of 0.02. The maximum number of epochs is 500. The model with the highest mean Intersection over Union (mIoU) on the validation set is adopted to explore the network's performance on the test set. In all cases, the Multiclass Cross Entropy (MCE) between prediction and ground truth is employed as the segmentation loss. Here, we adopt deep supervision for the four side-outputs (i.e., P_0, P_1, P_2, and P_3 in Fig. 4). Each map is upsampled to match the size of the GT. As a result, the total loss for CAENet can be formulated as: $Loss_{total} = \sum_i Loss_i = \sum_i \mathrm{MCE}(P_i, GT)$, $i \in \{0, 1, 2, 3\}$.

Evaluation criteria. We performed three-fold validation on both ablation and contrast experiments and finally reported the average results over three folds. To objectively assess the performance of the segmentation model, five quantitative evaluation metrics were employed, including mean Intersection over Union (mIoU), mean F1 score (mF1), Pixel Accuracy (PA), mean Recall (mRe), and Hausdorff Distance (HD). The specific calculations are shown in equations (12)–(16). In these equations, mIoU, mF1, PA, and mRe all indicate the overall similarity between the predicted results and the ground truth labels, where larger values indicate higher similarity. On the other hand, HD is related to the similarity at the edges of both the predicted results and the ground truth labels, with smaller values indicating higher edge similarity.

$$mIoU = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k}p_{ij} + \sum_{j=1}^{k}p_{ji} - p_{ii}} \tag{12}$$

$$mF1 = \frac{1}{k}\sum_{i=1}^{k}\frac{2 \times p_{ii}}{\sum_{j=1}^{k}p_{ij} + \sum_{j=1}^{k}p_{ji}} \tag{13}$$

$$PA = \frac{\sum_{i=1}^{k}p_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k}p_{ij}} \tag{14}$$

$$mRe = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{i=1}^{k}p_{ij}} \tag{15}$$

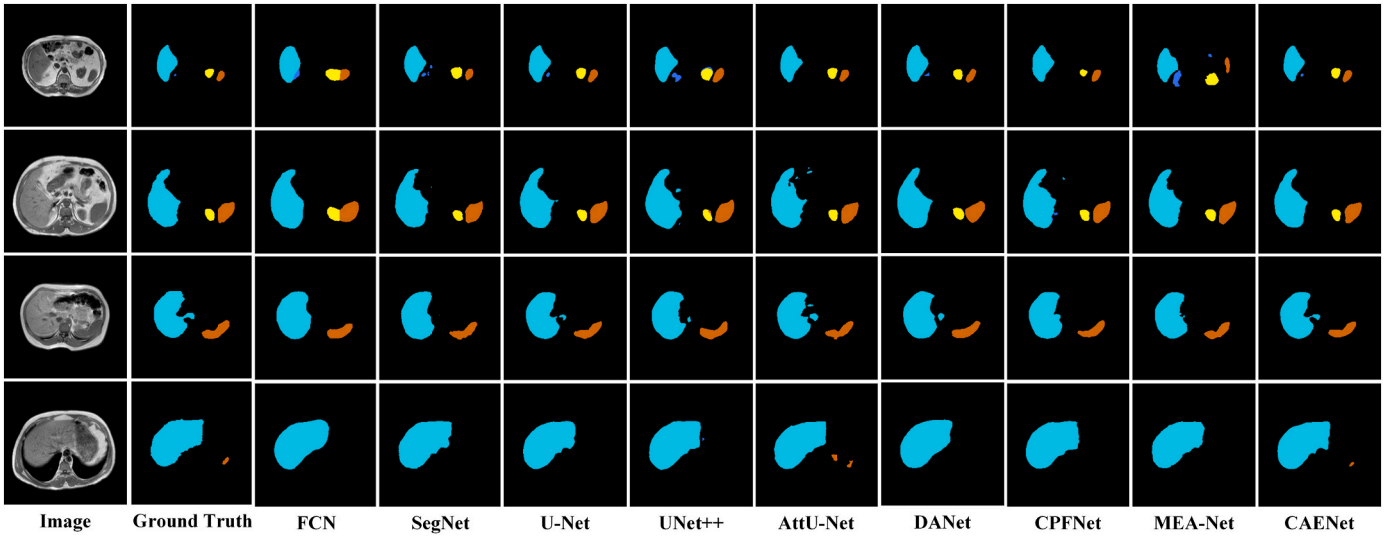

Fig. 5. Results for four subjects on the CHAOS-T1 dataset. The cyan, blue, yellow, and orange regions represent the liver, right kidney, left kidney, and spleen.

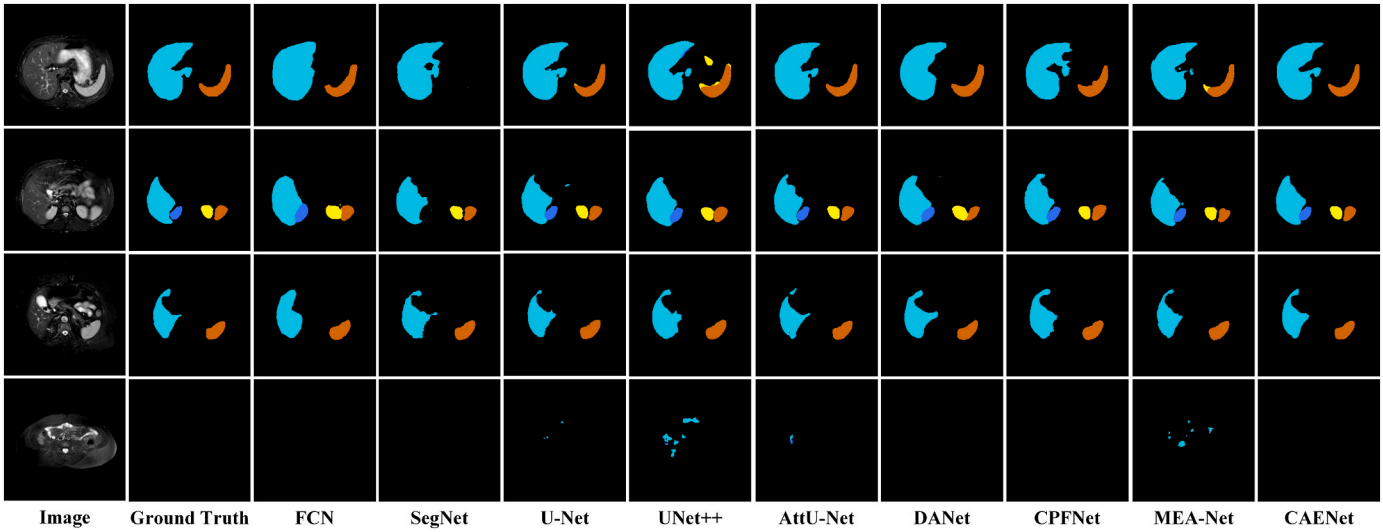

Fig. 6. Results for four subjects on the CHAOS-T2 dataset. The cyan, blue, yellow, and orange regions represent the liver, right kidney, left kidney, and spleen.

Table 3
Results of contrast experiments on the Lung dataset (mean ± standard deviation). The best results are shown in bold. The backbone of FCN and DANet are ResNet101 and ResNet50, respectively.

Network	mIoU (%)	mF1(%)	PA (%)	mRe (%)	HD (mm)
FCN [5]	94.98 ± 0.21	97.42 ± 0.11	98.80 ± 0.08	98.49 ± 0.04	6.35 ± 0.14
SegNet [8]	96.00 ± 0.21	97.96 ± 0.11	99.06 ± 0.07	98.31 ± 0.21	6.00 ± 0.14
U-Net [7]	96.87 ± 0.16	98.41 ± 0.08	99.27 ± 0.05	99.00 ± 0.18	5.65 ± 0.08
U-Net++ [9]	94.25 ± 1.99	97.03 ± 1.06	98.61 ± 0.51	98.53 ± 0.35	6.86 ± 0.89
AttU-Net [10]	96.46 ± 0.15	98.20 ± 0.08	99.17 ± 0.05	98.68 ± 0.20	5.94 ± 0.10
DANet [14]	95.23 ± 0.27	97.56 ± 0.14	98.86 ± 0.08	98.53 ± 0.00	6.34 ± 0.23
CE-Net [13]	96.21 ± 0.29	98.07 ± 0.15	99.10 ± 0.08	98.95 ± 0.09	6.01 ± 0.21
CPFNet [11]	96.11 ± 0.28	98.01 ± 0.15	99.08 ± 0.09	98.80 ± 0.11	5.92 ± 0.18
MEA-Net [16]	96.21 ± 0.21	98.07 ± 0.11	99.11 ± 0.06	98.80 ± 0.11	6.14 ± 0.03
CAENet (ours)	**97.54** ± **0.06**	**98.76** ± **0.03**	**99.42** ± **0.03**	98.92 ± 0.12	**5.31** ± **0.12**

Table 4
Results of contrast experiments on the Tongue dataset (mean ± standard deviation). The best results are shown in bold. The backbone of FCN and DANet are ResNet101 and ResNet50, respectively.

Network	mIoU (%)	mF1 (%)	PA (%)	mRe (%)	HD (mm)
FCN [5]	96.02 ± 0.21	97.97 ± 0.11	99.11 ± 0.02	98.30 ± 0.14	6.48 ± 0.10
SegNet [8]	96.42 ± 0.20	98.18 ± 0.10	99.20 ± 0.05	98.43 ± 0.20	6.25 ± 0.17
U-Net [7]	97.09 ± 0.16	98.52 ± 0.09	99.35 ± 0.04	98.73 ± 0.28	6.01 ± 0.05
U-Net++ [9]	94.93 ± 0.76	97.40 ± 0.40	98.86 ± 0.15	97.89 ± 0.55	6.51 ± 0.15
AttU-Net [10]	97.11 ± 0.14	98.54 ± 0.08	99.36 ± 0.02	98.68 ± 0.17	6.06 ± 0.10
DANet [14]	96.65 ± 0.15	98.30 ± 0.08	99.25 ± 0.04	98.76 ± 0.05	6.29 ± 0.10
CE-Net [13]	96.42 ± 0.11	98.18 ± 0.06	99.20 ± 0.02	98.62 ± 0.14	6.49 ± 0.11
CPFNet [11]	96.61 ± 0.15	98.27 ± 0.08	99.24 ± 0.03	98.63 ± 0.13	6.38 ± 0.12
MEA-Net [16]	96.42 ± 0.36	98.18 ± 0.18	99.21 ± 0.08	97.89 ± 0.47	6.29 ± 0.08
CAENet (ours)	**97.55** ± **0.07**	**98.76** ± **0.04**	**99.46** ± **0.01**	**98.88** ± **0.11**	**5.73** ± **0.09**

$$HD = \max\left(\max_{a\in A}\left\{\min_{b\in B}\|a-b\|\right\}, \max_{b\in B}\left\{\min_{a\in A}\|b-a\|\right\}\right) \tag{16}$$

Where k is the number of segmentation target classes, p_{ij} represents the number of pixels where true class i is predicted as class j, and $\|\cdot\|$ denotes the Euclidean distance between pixel sets A and B. In the context of our evaluation metrics, these calculations help quantify the degree of similarity between the predicted results and the ground truth labels for each class.

4.3. Segmentation results on CHAOS-T1 and CHAOS-T2 datasets

Quantitative comparative evaluation. Table 1 and Table 2 quantitatively show the comparison results between the proposed network CAENet and other excellent CNN methods on CHAOS-T1 and CHAOS-T2 datasets, respectively. CAENet outperforms the comparison networks in all evaluation criteria. In the results for the CHAOS-T1 dataset, CAENet achieves 88.21 %, 93.72 %, 99.55 %, 94.33 %, and 1.15 mm in mIoU, mF1, PA, mRe, and HD, respectively. Compared with several networks (i.e., U-Net, AttU-Net, CPFNet, and MEA-Net), CAENet realizes a mean improvement of 2.17 %, 1.48 %, 5.45 %, and 5.78 % for the main evaluation metric mIoU index, and 1.25 %, 0.85 %, 3.20 %, and 3.42 % for mF1 index, respectively, proving the effectiveness of CAENet in this dataset. For the CHAOS-T2 dataset (Table 2), compared with U-Net and AttU-Net with enhanced performance, CAENet improves by 3.65 % (on mIoU), 2.07 % (on mF1), 0.18 % (on PA), 2.38 % (on mRe), and 0.16 mm (on HD) for U-Net and 3.08 % (on mIoU), 1.74 % (on mF1), 0.18 % (on PA), 2.16 % (on mRe), and 0.15 mm (on HD) for AttU-Net. Experimental results show that the proposed network can be effectively applied to multiclass image segmentation and outperform the comparison methods.

Qualitative comparative evaluation. These samples (Fig. 5) from the CHAOS-T1 dataset pose various challenges, such as irregular objects and/or fuzzy details. According to Fig. 5, U-Net++ mainly suffers from category confusion and oversegmentation. For example, the blue target in the first row is partially identified as an orange target, and the cyan target in the third row and the background in the fourth row are oversegmented. The performance of CPFNet is inferior on small targets, such

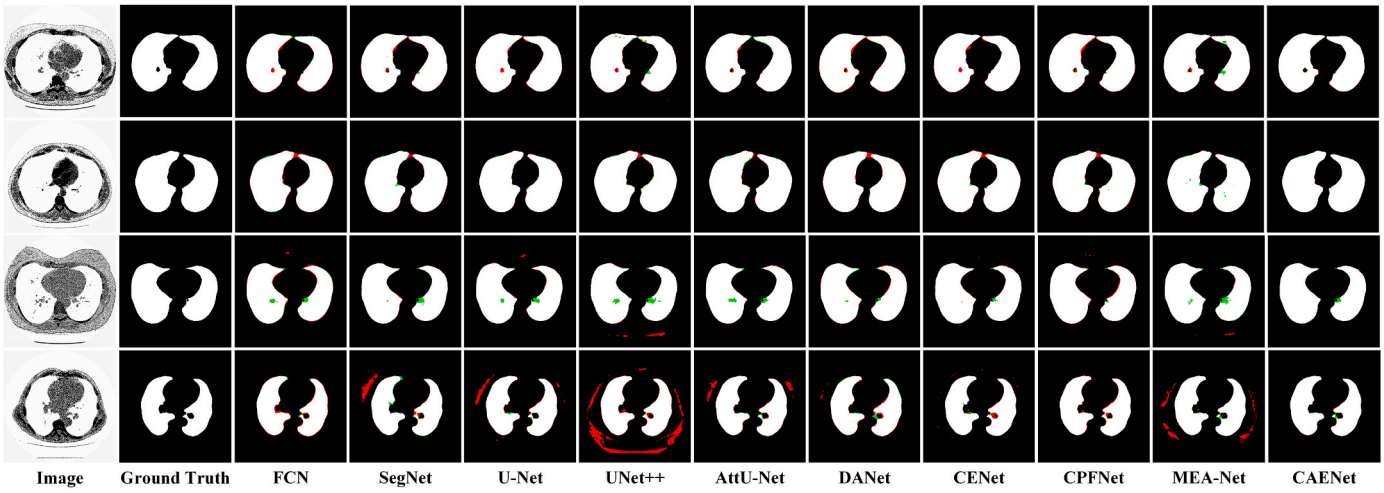

Fig. 7. Results of four subjects on the Lung dataset. The colors white, red, and green indicate correct segmentation, oversegmentation, and undersegmentation, respectively.

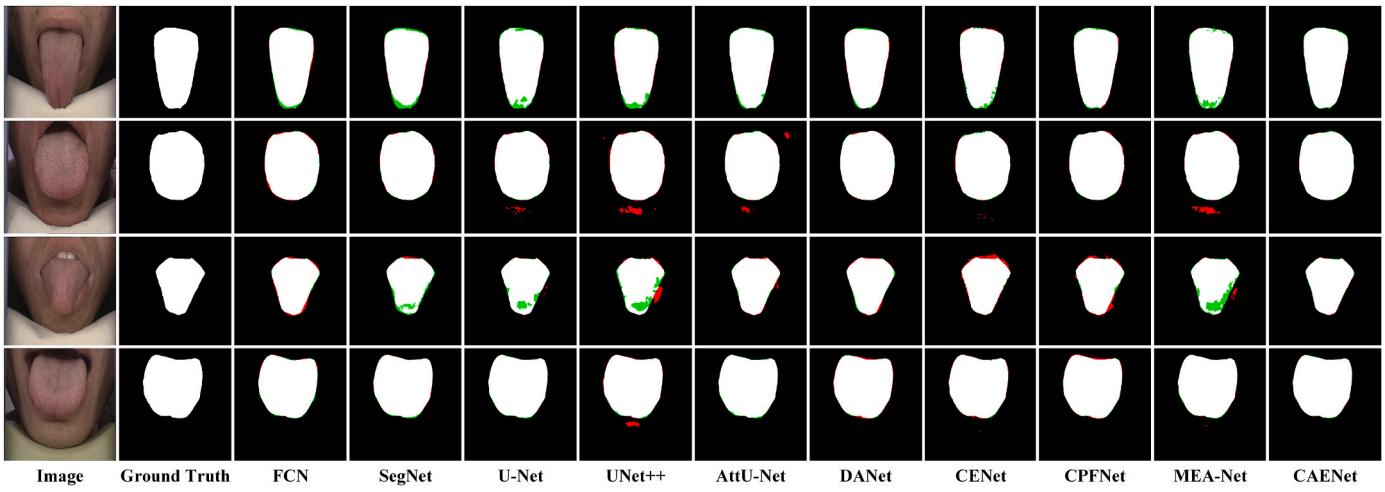

Fig. 8. Results of four subjects on the Tongue dataset. The colors white, red, and green indicate correct segmentation, oversegmentation, and undersegmentation, respectively. a. Results of four subjects on the Clinical-Face dataset. The white, yellow, cyan, gray, and orange regions represent the forehead-heart, left cheek-liver, nose-spleen, right cheek-lung, and jaw-kidney, respectively **b.** The effect of CAENet for foreground segmentation. The colors white, red, and green indicate correct segmentation, oversegmentation, and undersegmentation, respectively.

as the blue target in the first row and the orange target in the fourth row that are not identified. As can be seen in the fourth row of Fig. 5, the orange target is small and blurry, hindering segmentation. DANet, FCN, SegNet, U-Net, and MEA-Net failed to correctly identify it; although AttU-Net segmented it, it was oversegmented. The proposed CAENet can effectively identify small targets and segment more details, and the segmentation results are closer to the true labels.

As can be seen in Fig. 6, the CHAOS-T2 dataset suffers from coarse details and class confusion. Among the compared methods, although FCN, DANet, and CPFNet exhibit suitable category recognition capability, they provide coarse segmentation details. SegNet exhibits an inferior ability to recognize different classes, and the identified edges of large objects are coarse and uneven. Although U-Net++ reduces the semantic gap, it may introduce noisy information, which leads to category confusion, oversegmentation, and insufficient detail discrimination in segmentation results. MEA-Net exhibits a certain ability to recognize edges; however, the recognition of details is not sufficiently accurate, and the bit category confusion and background misjudgment indicate the insufficiency of the feature extraction ability of MEA-Net on this dataset. U-Net and AttU-Net are more accurate in shape information recognition; nonetheless, background misidentify exists (e.g., Row 4). The proposed CAENet exhibits superior semantic category recognition ability and can retain more valuable details, such as the edge of the cyan category in the first and second rows, the gap between adjacent categories in the second row, the connectivity of the cyan category in the third row, and the correct recognition of background in the fourth row.

4.4. Segmentation results on LUNA and tongue datasets

Quantitative comparative evaluation. Table 3 and Table 4 quantitatively show the comparison results between the proposed network CAENet and other excellent methods on Lung and Tongue, respectively, including FCN, SegNet, U-Net, U-Net++, AttU-Net, DANet, CE-Net, CPFNet, and MEA-Net. Overall, the performance of each network on the two datasets is relatively efficient, and the segmentation performance of each one is above 94 % (on mIoU), among which CAENet achieves the optimal results on five comparison indexes.

Quantitative comparative evaluation. Images from Lung and Tongue datasets are both single-target segmentation tasks. The difference is that the Lung dataset is a single-target, multiregion segmentation

Table 5
Segmentation performances of different networks on Clinical-Face dataset (mean ± standard deviation). The best results are shown in bold. The backbone of FCN and DANet are ResNet101 and ResNet50, respectively.

Network	mIoU (%)	mF1 (%)	PA (%)	mRe (%)	HD (mm)
FCN [5]	47.43 ± 11.32	63.19 ± 11.33	96.49 ± 1.11	57.13 ± 11.92	4.25 ± 0.45
SegNet [8]	31.60 ± 12.04	81.56 ± 2.62	96.30 ± 0.88	37.12 ± 13.97	4.88 ± 0.48
U-Net [7]	78.71 ± 0.45	88.06 ± 0.28	98.65 ± 0.10	88.32 ± 0.88	3.12 ± 0.10
U-Net++ [9]	56.75 ± 5.21	72.11 ± 4.41	96.54 ± 0.72	85.54 ± 4.38	3.99 ± 0.30
AttU-Net [10]	72.07 ± 1.26	83.67 ± 0.84	98.25 ± 0.16	81.52 ± 1.17	3.33 ± 0.07
DANet [14]	34.23 ± 7.73	50.20 ± 8.83	92.45 ± 1.69	66.44 ± 2.56	4.98 ± 0.82
CPFNet [11]	35.31 ± 7.10	45.73 ± 14.21	96.07 ± 0.69	38.14 ± 9.27	4.56 ± 0.28
MEA-Net [16]	70.79 ± 1.82	82.83 ± 1.29	98.20 ± 0.25	78.64 ± 2.54	3.40 ± 0.13
CAENet (ours)	**80.82 ± 0.67**	**89.38 ± 0.40**	**98.76 ± 0.17**	**89.13 ± 1.06**	**2.95 ± 0.10**

task, whereas the Tongue dataset is a single-target, fully connected region segmentation task. The difficulty of segmentation in Lung images mainly lies in the accurate segmentation of edges and complex connected regions, which in Tongue images mainly lies in the accurate differentiation of tongue and lip colors and the accurate identification of irregular edges. From Figs. 7 and 8, the compared networks frequently result in oversegmentation or undersegmentation. In comparison, the qualitative results of CAENet in Figs. 7 and 8 show that the network is more dominant in terms of edge and tongue segmentation.

4.5. Segmentation results on Clinical-Face dataset

Quantitative comparative evaluation. To validate the generalization performance of the proposed CAENet, we conducted experiments on the Clinical-Face dataset, a clinical TCM-related dataset. The results in Table 5 present that CAENet reaches better performance than all the comparative networks. Unlike U-Net, it exhibits an overall improvement (2.11 % for the mIoU index, 1.32 % for the mF1 index, 0.11 % for the PA index, 0.81 % for the mRe index, and 0.17 mm for the HD index). Since the data set is collected in an open environment, the background of the image is complex and the pixel value is rich, which makes the segmentation more difficult. Therefore, each network in the performance of the dataset partition is relatively lacking and the proposed CAENet network on the various indicators of optimal goal proves the method's complex background processing ability.

Qualitative comparative evaluation. Images in the Clinical-Face datasets captured in an open environment are vulnerable to light intensity and complex backgrounds, which can hinder segmentation. Fig. 9a shows that U-Net and AttU-Net make relatively accurate predictions; however, a few oversegmented and undersegmented edges exist. U-Net++ is insensitive to intercategory interval information, resulting in category confusion. Each category of CPFNet is identified;

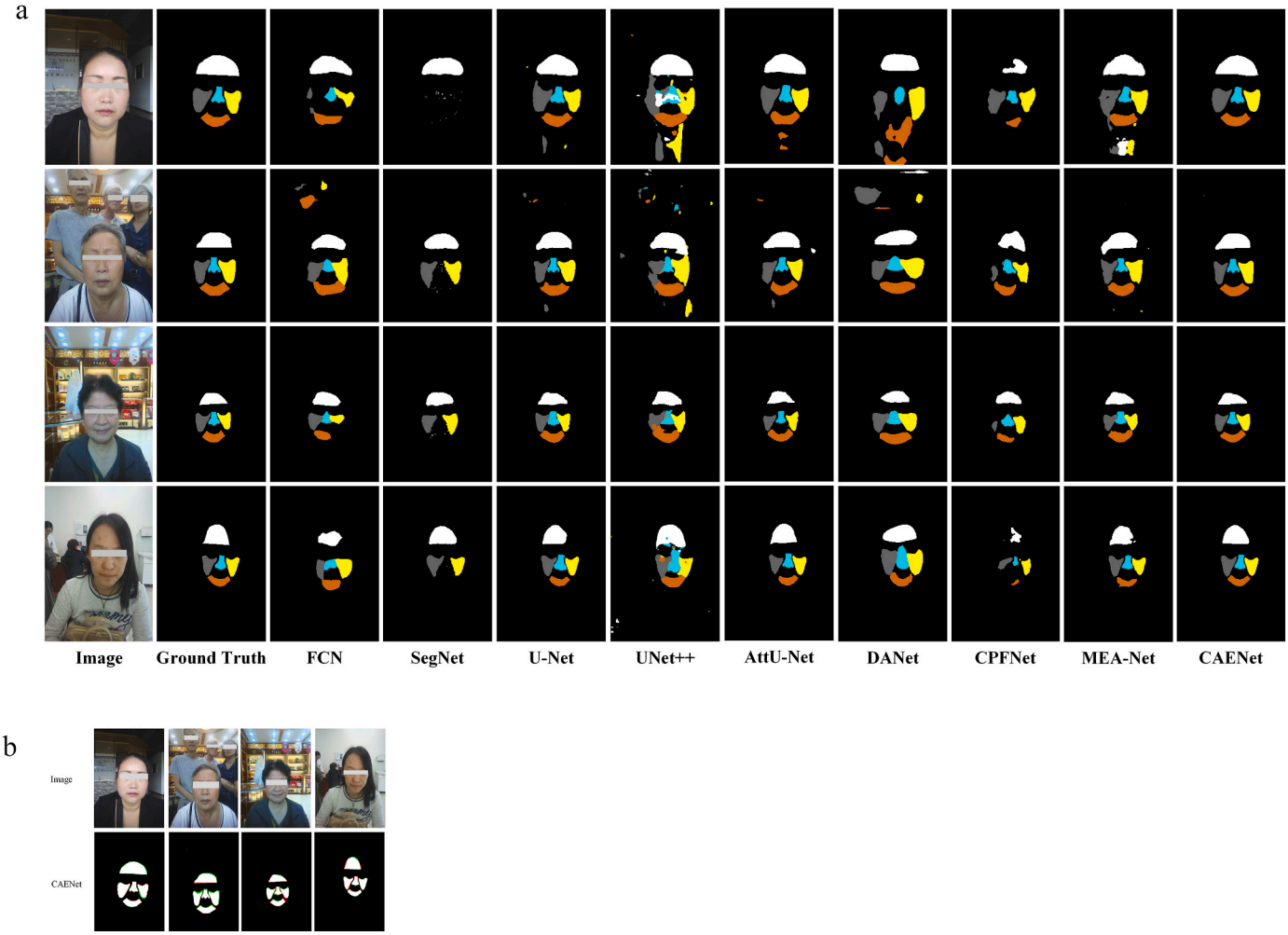

Fig. 9. Results of four subjects on the Clinical-Face dataset.

Table 6
Ablation study for different modules on CHAOS-T1 and CHAOS-T2 datasets (mean ± standard deviation). The best results are shown in bold.

Dataset	Method	mIoU (%)	mF1 (%)	PA (%)	mRe (%)	HD (mm)
CHAOS-T1	Baseline	86.77 ± 0.56	92.90 ± 0.32	99.50 ± 0.06	93.71 ± 1.20	1.21 ± 0.09
	Model 1	87.57 ± 0.47	93.36 ± 0.27	99.53 ± 0.04	**94.50 ± 0.10**	1.18 ± 0.08
	Model 2	87.13 ± 0.98	93.10 ± 0.56	99.51 ± 0.05	94.01 ± 0.88	1.22 ± 0.06
	Model 3	86.97 ± 0.81	93.01 ± 0.47	99.51 ± 0.05	92.79 ± 1.20	1.19 ± 0.09
	Model 4	87.52 ± 0.90	93.33 ± 0.51	99.52 ± 0.05	93.93 ± 1.17	1.19 ± 0.05
	Model 5	86.84 ± 1.07	92.93 ± 0.62	99.50 ± 0.06	94.14 ± 0.13	1.22 ± 0.06
	Model 6	87.24 ± 0.74	93.16 ± 0.42	99.52 ± 0.04	93.80 ± 0.99	1.22 ± 0.06
	Model 7	87.57 ± 0.88	93.35 ± 0.50	99.53 ± 0.05	93.77 ± 0.99	1.18 ± 0.06
	Model 8	87.67 ± 1.05	93.41 ± 0.60	99.53 ± 0.05	93.79 ± 1.45	1.18 ± 0.08
	CAENet	**88.21 ± 0.38**	**93.72 ± 0.21**	**99.55 ± 0.04**	93.99 ± 0.14	**1.15 ± 0.09**
CHAOS-T2	Baseline	87.12 ± 2.01	93.03 ± 1.17	99.34 ± 0.06	91.83 ± 1.74	1.42 ± 0.08
	Model 1	88.81 ± 1.53	94.06 ± 0.87	99.44 ± 0.04	92.97 ± 1.84	1.29 ± 0.04
	Model 2	87.08 ± 1.99	93.06 ± 1.16	99.32 ± 0.06	91.85 ± 2.07	1.44 ± 0.05
	Model 3	87.63 ± 1.12	93.40 ± 0.64	99.35 ± 0.03	92.32 ± 1.08	1.41 ± 0.03
	Model 4	88.15 ± 1.39	93.69 ± 0.8	99.40 ± 0.04	92.85 ± 0.98	1.37 ± 0.03
	Model 5	87.70 ± 1.64	93.43 ± 0.94	99.36 ± 0.06	93.67 ± 0.38	1.39 ± 0.03
	Model 6	87.78 ± 0.73	93.48 ± 0.42	99.38 ± 0.02	92.83 ± 0.52	1.37 ± 0.03
	Model 7	88.44 ± 0.84	93.85 ± 0.48	99.41 ± 0.03	93.44 ± 0.49	1.35 ± 0.02
	Model 8	88.49 ± 1.22	93.88 ± 0.69	99.39 ± 0.03	92.97 ± 1.84	1.38 ± 0.04
	CAENet	**89.95 ± 0.86**	**94.70 ± 0.48**	**99.49 ± 0.04**	**94.33 ± 1.10**	**1.26 ± 0.03**

nonetheless, undersegmentation is severe. SegNet only segments regions with large shapes and sharp edges. DANet incorrectly identified the neck as the jaw in the first row and the nontargeted face region in the background as the target in the second row, indicating its weak antijamming capability. The classification accuracy of FCN is higher than that of CPFNet; however, it loses extensive details and is insensitive to target location and edge information. The proposed CAENet is still able to accurately localize the target and accurately identify the category information in diverse backgrounds, and the edge segmentation effect is closer to the true label, indicating that the network is robust against interference and suitable for complex background segmentation tasks. Fig. 9b shows the oversegmentation and undersegmentation of CAENet for targets, indicating that the main deficiency lies in the edge regions. These deficiencies can be further strengthened and improved in future work.

5. Ablation experiment

To evaluate the individual contribution of different components of our proposed network to the segmentation performance, we also performed a stepwise ablation experiment on the CHAOS-T1 and CHAOS-T2 datasets. The models compared are as follows:

Baseline: Consider the basic U-shape model shown in Fig. 1 as a baseline method.

Model 1: Baseline + DeepSup.

Model 2: Baseline + SAM.

Model 3: Baseline + ACA.

Model 4: Baseline + SMPool1.

Model 5: Baseline + FECA.

Model 6: Baseline + FECA + SAM.

Model 7: Baseline + ACA + SMPool1.

Model 8: Baseline + ACA + SMPool1+FECA + SAM.

CAENet (ours): Baseline + ACA + SMPool1+FECA + SAM + DeepSup.

Table 6 provides detailed quantitative experimental results for Baseline and Baseline with multiple module compositions. The addition of DeepSup, ACA module, and SMPool1 pooling function, respectively, improve over the Baseline (i.e., the mIoU index increases 0.8 %, 0.20 %, and 0.75 % for CHAOS-T1, 1.69 %, 0.51 %, and 1.03 % on the CHAOS-T2 dataset), indicating that each of these methods positively influences the Baseline. The gains obtained by DeepSup and SMPool1 over Baseline clearly show that the auxiliary optimization at the decoding end of each layer enables improving the overall performance of the model and that max pooling leads to information loss during downsampling. In addition, the differentiable approximate pooling function SMPool1 enhances the connections between neurons in the network, such that more useful information can be retained during encoding.

Compared with SMPool1 and ACA alone, the combination of SMPool1 and ACA is superior, indicating that the combination of SMPool1 and ACA is more effective than the single one. FECA emphasizes the importance of channel features, while SAM filters out spatial details. As can be seen from Table 6, adding FECA and SAM alone does not always improve the performance of the model, as the attention may not make more accurate judgments when it does not extract more useful information. The ACA module combined with SMPool1 enhances high-frequency information and ensures continuity of information transmission; however, it may transmit redundant information simultaneously. Therefore, FECA and SAM are added to screen important features and achieve better results in all metrics. Finally, the output auxiliary training of each decoding layer is supervised by DeepSup to optimize the final output result. The final segmentation performance on the CHAOS-T1 dataset outperforms the Baseline by 1.44 % (on mIoU), 0.82 % (on mF1), 0.05 % (on PA), 0.28 % (on mRe), and 0.06 mm (on HD), respectively. For the CHAOS-T2 dataset, in contrast to Baseline, the proposed method improves by 2.83 % (on mIoU), 1.67 % (on mF1), 0.15 % (on PA), and 0.16 mm (on HD).

6. Discussion

6.1. ACA module study

Fig. 10 illustrates the augmented feature maps obtained via different contrast augmentation methods, where the pixel values are represented by different colors, with higher colors being warmer. As can be seen from Fig. 10, the enhancement of LCS and HE is for the global image, where LCS enhances the background more than HE. Both enhance the edge information and the high-frequency noise within the class, such as the distinction between the details of the cyan category in lines 1–2, causing loss of category information and rendering it adverse to subsequent feature extraction and semantic category recognition. The CLAHE algorithm is a limited local contrast enhancement method, which is based on setting a threshold on the gray level of the image histogram, clipping the excess part, and assigning it equally to each gray level. According to the results in Fig. 10, the enhanced feature maps perform better than LCS and HE, and the high-frequency information is enhanced while the integrity of the target is mostly preserved.

However, the enhancement effect of this method varies for different input images. For example, in the second row, over-enhancement within the class and background still exists. This is because the threshold of the CLAHE method needs to be set manually; different input images are suitable for different thresholds, and the effects of different images at the

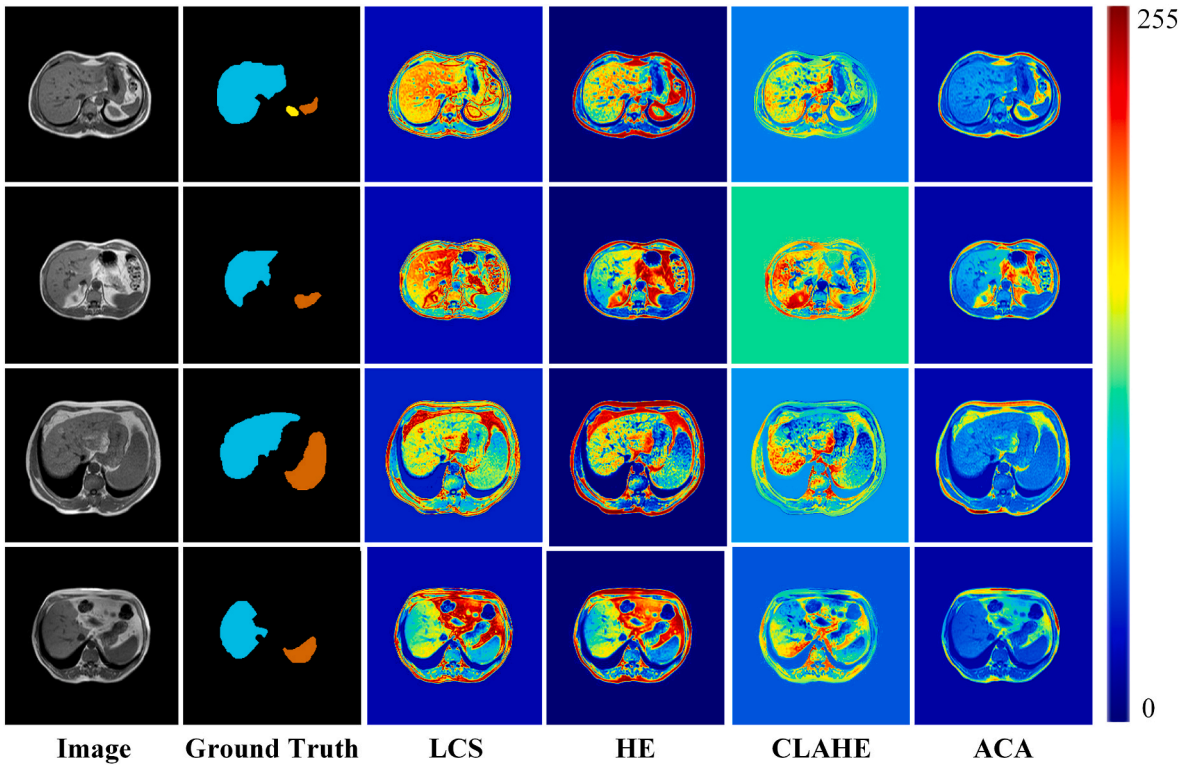

Fig. 10. Visualization results of the different methods for image augmentation. The warmer color indicates a higher value. The color band is listed at the rightmost end of the image.

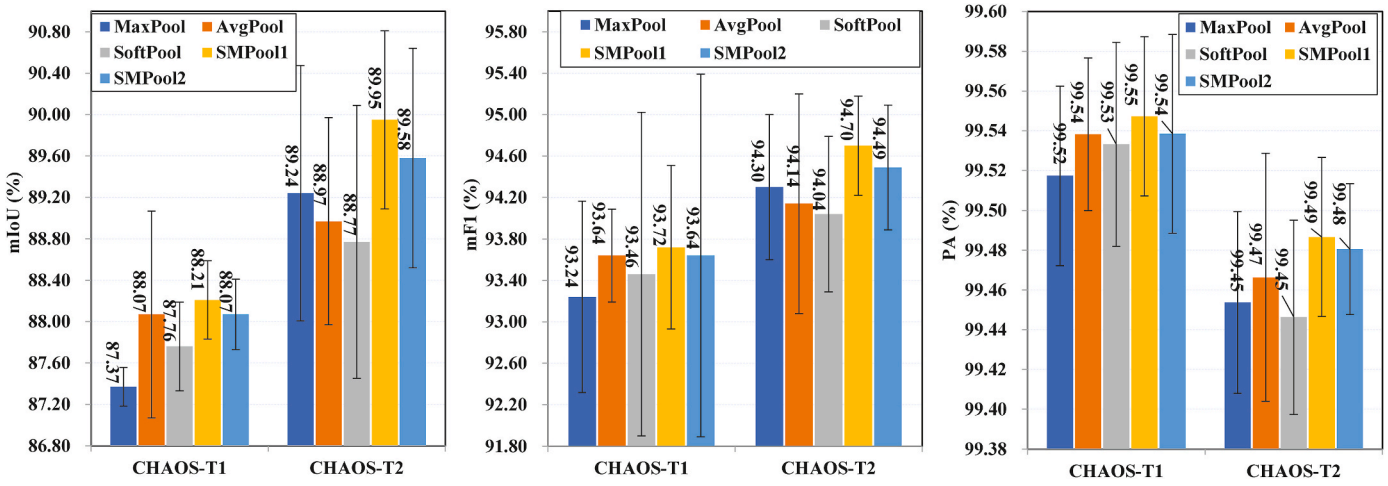

Fig. 11. Segmentation performance of CAENet with different pooling functions.

Table 7
Statistics of 10 pooling replacements on different networks (mean ± standard deviation).

model	Before		After		P-value	
	mIoU (%)	mF1 (%)	mIoU (%)	mF1 (%)	mIoU	mF1
U-Net	84.13 ± 0.53	91.35 ± 0.32	85.13 ± 0.64	91.94 ± 0.38	1.32e-3	1.52e-3
AttU-Net	85.28 ± 054	92.03 ± 0.32	86.14 ± 068	92.54 ± 0.40	5.83e-3	6.08e-3
MEA-Net	80.90 ± 0.56	89.36 ± 0.35	82.42 ± 0.51	90.28 ± 0.32	6.00e-6	9.00e-6
CAENet	87.59 ± 0.20	93.37 ± 0.12	88.60 ± 0.43	93.95 ± 0.24	1.40e-5	1.40e-5

same threshold are different. The proposed ACA module focuses on target boundaries while avoiding the over-enhancement of classes. This is because the method adjusts the enhancement scope automatically during the training process, and enhances the edge between categories and the background between categories, which on reservation of target information simultaneously increases the degree of differentiation between target and nontargeted classes, reducing the introduction of high-frequency noise.

6.2. *Pooling method study*

Comparison of the effect of different pooling functions. Fig. 11 displays the segmentation performance of CAENet with different pooling methods on the two datasets. According to the results, the proposed

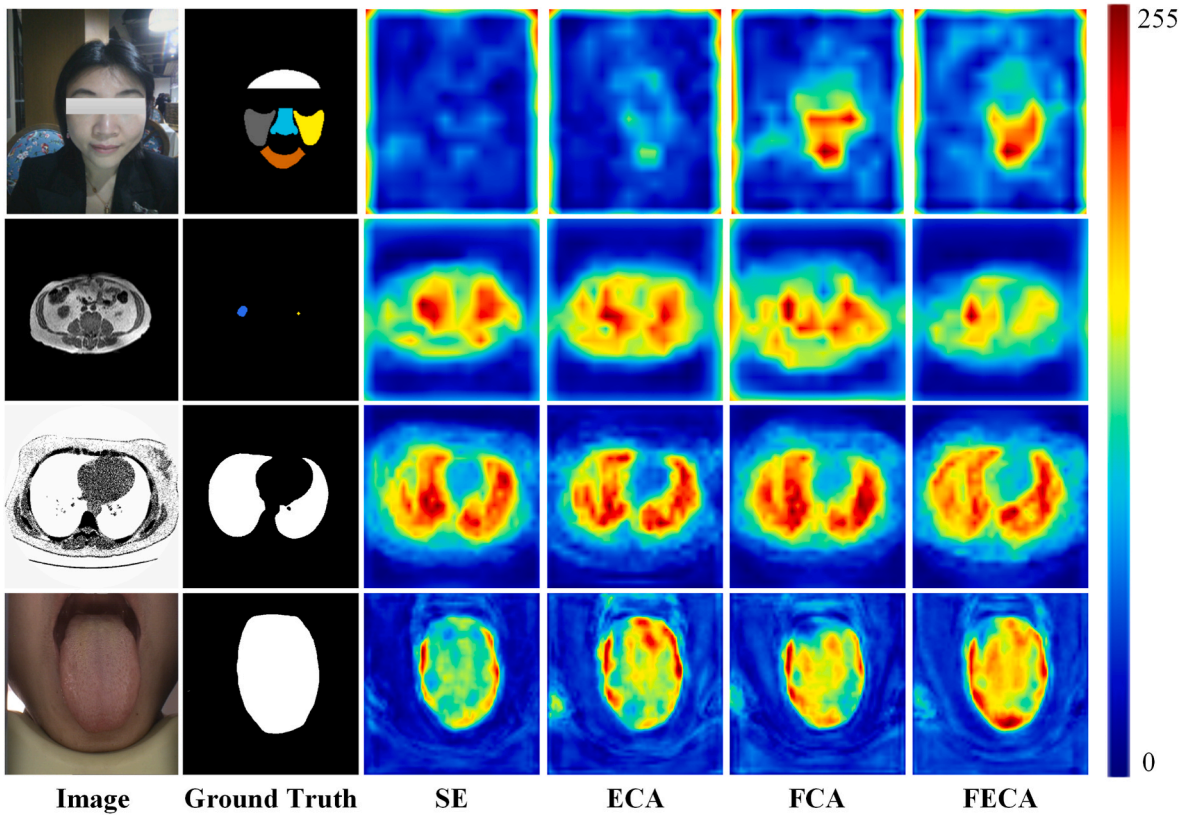

Fig. 12. Channel maps arise from different attention mechanisms. The warmer color indicates a higher value. The color band is listed at the rightmost end of the image.

Table 8
The parameter increments introduced by different attention methods.

Methods	Params (K)
SE	76.288
ECA	0.023
FCA	76.288
FECA(ours)	**0.015**

differentiable pooling function can enable improving the performance of CAENet, and SMPool1 achieves the best results compared with the others on three main metrics. Compared with MaxPool, AvgPool, SMPool1, and SMPool2 perform better, indicating that global information can retain more useful information than local maxima during downsampling. Among them, AvgPool achieves similar performance as SMPool1 in terms of index results. However, Fig. 11 demonstrates that AvgPool has a larger standard deviation, this is attributed to the fact that average pooling tends to dilute the impact of all activations within the pooling region, especially when there are multiple crucial features present in the pooling window. In contrast, SMPool mitigates this information loss, enhancing model stability and robustness. Furthermore, the varying performance of SoftPool suggests weaker generalization ability on different datasets. Similarly, SMPool1 functions as a nonlinear operation where larger activation values contribute more to the output. However, SMPool1's computation is simpler, and the output results are regulated by the trainable parameter β. As the values of β adjust and change across different pooling layers due to training, SMPool becomes more adaptable to data variations, ultimately demonstrating improved performance and stability.

SMPool replaces the non-differentiable max pooling operation, enabling even slight variations to result in corresponding output changes. For each feature map, the value of β is determined, with larger activation values exerting a stronger influence on the output. This indicates that SMPool ensures the representation of prominent features while also considering the contributions of other features, delivering more comprehensive information. For MaxPool, the mIoU index on CHAOS-T1 and CHAOS-T2 datasets are 87.37 % and 89.24 %, respectively, whereas the mIoU index is 88.21 % (an increase of 0.84 %) and 89.95 % (an increase of 0.71 %) for SMPool. When using SMPool2, CAENet achieves suboptimal results on CHAOS-T1 and CHAOS-T2 datasets, affording the mIoU index of 88.07 % (an increase of 0.77 %) and 89.58 % (an increase of 0.34 %). As can be seen from the overall results, SMPool1 and SMPool2 achieve optimal and suboptimal results, respectively, in terms of the three main metrics—mIoU, mF1, and PA—for both datasets.

The substitution effect of pooling function on different networks. In order to validate the generalization performance of the pooling function on different models, we further performed 10 experiments with the pooling function replacement on the first fold data from the public dataset CHAOS-T2 independently. The statistical results are shown in Table 7, the compared networks show stable improvement in mIoU and mF1 after replacing the original pooling function with SMPool1. The results show that the proposed pooling function is also effective in improving the performance of other segmentation networks, with some generalization.

The statistical analysis software SPSS was used to test the normality of the 10 statistical results for the two metrics for each network, and the nonparametric test results showed that the statistical results all satisfy the normality. In addition, the average values of the results before and after pooling replacement for different models are compared. Student t-test was performed for those that met the homogeneity of variance and the Satterthwaite *t*-test was performed for those that did not meet the requirement. Table 7 presents that all values of P-value are much smaller

than 0.05, which corresponds to confidence 95 % that the improved results are indeed due to the replaced pooling. In other words, the improvement caused by the replacement of the pooling function is statistically significant.

6.3. Attention mechanism study

To understand the effect of the attention mechanism more clearly, we visualized the semantic heat maps of channel outputs after attention modules at the encoding end, as shown in Fig. 12.

The response of a specific semantic category is more noticeable after channel attention modules. Although SE, ECA, and FCA can also highlight specific class semantics, there are still certain nontargeted regions in the semantic maps of small targets, such as small targets in the second row that cannot be appropriately distinguished in the corresponding attentive feature maps. The proposed FECA module generates a feature (the sixth column) that better focuses on the specific regions of the structures of interest and simultaneously focuses on the size and shape of the region, avoiding ambiguous regions that might result in misclassification.

In contrast, FECA excels in precise target identification by utilizing 2DDDCT to calculate multi-frequency information. It focuses on the highest-energy low-frequency, akin to the effect of average pooling, while also considering information carried by other frequency components. This approach enhances the model's perception of distinct frequency features, resulting in improved performance. The results in the first row of Fig. 12 illustrate that both FECA and FCA successfully identify blurred semantic objects, whereas SE and ECA fall short in this regard. Comparing ECA and SE, they yield similar outcomes, with ECA being more lightweight in terms of parameters (Table 8). Notably, FECA achieves an even lighter parameter count, introducing only 0.015K parameters, while delivering more accurate performance.

7. Conclusion

In this paper, we have proposed a novel CAENet network with the ability to efficiently identify details and small targets. First, the ACA module is designed to enhance the contrast of the input images, while the approximation of differentiable max pooling is introduced to retain more useful information during the downsampling phase. In addition, the channel attention FECA is proposed to counter the introduction of redundant channel feature information. Furthermore, the fully connected layer is replaced by a one-dimensional convolution with convolution kernel size 3 based on multifrequency information to maintain accuracy and reduce computational complexity. Finally, the segmentation performance of CAENet is validated by ablation and contrast experiments on five medical image datasets. Experimental results show that the proposed CAENet identifies small objects and details more effectively and achieves the best comparison results in five quantitative metrics, indicating that the network exhibits strong generalization ability in medical image segmentation and can be further applied to other medical image segmentation tasks.

The modules presented in this paper can be easily ported into other networks. However, the convolution of the encoding and decoding layers is not changed, which may limit the segmentation performance. Therefore, in future studies, we will explore new encoder and decoder structures to build a more efficient segmentation network in combination with the approach proposed in this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province, China (2021ZDZX1032); the Special Project of Guangdong Province, China (2020A1313030021); and the Scientific Research Project of Wuyi University, China (2018TP023, 2018GR003). (Shengke Li and Yue Feng contributed equally to this article. The authors would like to thank Xin Wu for her excellent technical support.)

We extend our gratitude to all colleagues who provided valuable suggestions and support throughout this research. We are open to sharing our implementation code and additional details with researchers interested in our study. For access to the code or further information, please contact the corresponding author.

References

[1] S. Wang, et al., Annotation-efficient deep learning for automatic medical image segmentation, Nat. Commun. 12 (1) (2021) 1–13.
[2] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[3] M. Antonelli, et al., The medical segmentation decathlon (in English), Nat. Commun. 13 (1) (Jul 2022) 13. Art. no. 4128.
[4] S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET), IEEE, 2017, pp. 1–6.
[5] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (Apr 2017) 640–651.
[6] S. Qiu, C. Li, Y. Feng, S. Zuo, H. Liang, A. Xu, GFANet: gated fusion attention network for skin lesion segmentation, Comput. Biol. Med. 155 (2023), 106462.
[7] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, Springer, Cham, 2015, pp. 234–241, 2015.
[8] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[9] Z.W. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J.M. Liang, UNet plus plus : a nested U-net architecture for medical image segmentation, in: 4th International Workshop on Deep Learning in Medical Image Analysis (DLMIA)/8th International Workshop on Multimodal Learning for Clinical Decision Support (ML-CDS), Granada, SPAIN vol. 11045, Springer International Publishing Ag, CHAM, 2018, pp. 3–11, 2018.
[10] J. Schlemper, et al., Attention gated networks: learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207.
[11] S. Feng, et al., CPFNet: context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imag. 39 (10) (2020) 3008–3018.
[12] R. Gu, et al., CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imag. 40 (2) (2020) 699–711.
[13] Z. Gu, et al., Ce-net: context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imag. 38 (10) (2019) 2281–2292.
[14] J. Fu, et al., Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, IEEE Computer Society, Los Alamitos, 2019, pp. 3141–3149, 2019.
[15] A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation, Ieee Journal of Biomedical and Health Informatics 25 (1) (Jan 2021) 121–130 (in English).
[16] H. Liu, et al., MEA-Net: multilayer edge attention network for medical image segmentation, Sci. Rep. 12 (1) (2022) 1–15.
[17] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: Making vgg-style convnets great again, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, IEEE Computer Society, Los Alamitos, 2021, 2021, pp. 13733–13742.
[18] D. Dai, et al., Ms RED: a novel multi-scale residual encoding and decoding network for skin lesion segmentation, Med. Image Anal. 75 (2022), 102293.
[19] Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: frequency channel attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, IEEE, Piscataway, 2021, pp. 783–792, 2021.
[20] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023.
[21] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: efficient channel attention for deep convolutional neural networks, in: Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, IEEE Press, Piscataway, 2020, pp. 11531–11539, 2020.
[22] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), Cham, Springer, Cham, 2018, pp. 3–19, 2018.
[23] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, IEEE, USA, 2021, pp. 13708–13717.

[24] M.-H. Guo, et al., Attention Mechanisms in Computer Vision: A Survey, Computational Visual Media, 2022, pp. 1–38.
[25] Y. Zhou, C. Shi, B. Lai, G. Jimenez, Contrast enhancement of medical images using a new version of the world cup optimization algorithm, Quant. Imag. Med. Surg. 9 (9) (2019) 1528–1547.
[26] N.H. Kaplan, I. Erer, D. Kumlu, Image enhancement methods for remote sensing: a survey, in: *Remote Sensing*: IntechOpen, 2021.
[27] A. Stergiou, R. Poppe, G. Kalliatakis, Refining activation downsampling with SoftPool, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, IEEE, Montreal, 2021, pp. 10357–10366, 2021.
[28] D. Yu, H. Wang, P. Chen, Z. Wei, Mixed pooling for convolutional neural networks, in: International Conference on Rough Sets and Knowledge Technology, Shanghai, China vol. 8818, Springer, Cham, 2014, pp. 364–375, 2014.
[29] C.-Y. Lee, P. Gallagher, Z. Tu, Generalizing pooling functions in CNNs: mixed, gated, and tree, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 863–875.
[30] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-Net: going deeper with nested U-structure for salient object detection, Pattern Recogn. 106 (2020), 107404.
[31] D.-P. Fan, et al., Pranet: parallel reverse attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 263–273.
[32] T.-C. Nguyen, T.-P. Nguyen, G.-H. Diep, A.-H. Tran-Dinh, T.V. Nguyen, M.-T. Tran, CCBANet: cascading context and balancing attention for polyp segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, Springer, 2021, pp. 633–643. September 27–October 1, 2021, Proceedings, Part I 24.
[33] G. Kreisselmeier, R. Steinhauser, Systematic control design by optimizing a vector performance index, in: M.A. Cuenod (Ed.), Computer Aided Design of Control Systems, Pergamon, 1980, pp. 113–117.
[34] A. Kavur, M. Selver, O. Dicle, M. Barış, N. Gezer, in: Zenodo (Ed.), CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, 2019.
[35] A.E. Kavur, et al., CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation, Med. Image Anal. 69 (2021), 101950.
[36] A.E. Kavur, et al., Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors, Diagn. Interventional Radiol. 26 (1) (Jan 2020) 11–21.