# CAAGP: Rethinking Channel Attention with Adaptive Global Pooling for Liver Tumor Segmentation

Chi Zhang Jingben Lu Luxi Yang Chunguo Li*
School of Information Science and Engineering, Southeast University
{zhang_chi, lujingben, lxyang, chunguoli}@seu.edu.cn

## Abstract

*Channel attention, a channel-wise method often used in computer vision tasks, including liver tumor segmentation task, can model the channel relationship to augment the representation ability of feature maps, adaptively generating channel-wise responses using global pooling, which aggregates spatial information roughly. Actually, global pooling tends to lose more fine information, which is vital for segmentation tasks, consequently contributing to suboptimal performance. Hence, we rethink the problem and propose the channel attention with adaptive global pooling(short for CAAGP), which preserves spatial and fine-grained information for liver tumor segmentation tasks when generating channel attention. The model consists of three main parts, i.e., improved self-attention, adaptive global pooling and responses generation modules. With respect to the computing of the spatial attention, self-attention has achieved almost the most optimal performance, however introducing serious calculation and memory burdens. Therefore, we improve self-attention and consider aggregating spatial information from x and y directions respectively, effectively absorbing spatial information with little calculation cost. Extensive experiments have been conducted to verify the effectiveness of our proposed method and our CAAGP outperformed the channel attention with naive global pooling and other algorithms significantly in liver tumor segmentation, especially for small tumors.*

***Keywords:*** *channel attention, self-attention, adaptive global pooling, liver tumor segmentation*

## 1. Introduction

Convolutional neural networks [29, 30, 39, 13, 22] have been proven to be useful models for computer vision tasks, including liver tumor segmentation task. Recently, many attention mechanisms [12, 36, 23, 35, 7, 15] have been applied to the CNNs to improve the networks performance, however applications of which in liver tumor segmentation are not mature enough. To the best of our knowledge, the most popular attention mechanisms can be broadly divided into three categories, i.e., channel attention [12], naive spatial attention [36, 23] and self-attention [35], all of which are shown in Figure 2 (a), (b), (c) and (d) respectively. Theoretically speaking, despite that self-attention [35] is a spatial attention mechanism based on spatial information modeling, we still treat it separately from other spatial attention mechanisms, because of the different modeling approaches for long-range spatial attention. Self-attention [35] is global while the naive spatial attention [36, 23] is local.

Channel attention is one of the typical methods widely used in computer vision tasks, and is common in biomedical image segmentation task [21, 20, 3], which generates channel-wise responses adaptively, applied to the features from different channels, enhancing the representation ability of feature maps. For liver tumor segmentation tasks, the size of target tumor is often too small with no obvious features to detect, which usually needs abundant fine-grained information. However, the calculation of channel-wise response is considered too coarse for its design of spatial dimensions squeezing [12], with global average pooling [18], causing a great deal of information loss. Channel attention performs badly when handling fine-grained information of liver tumor. Former methods, e.g., SE-Net [12], pay little attention to this problem. We rethink the question carefully, and propose adaptive global pooling when conducting channel attention in segmentation tasks, termed as Channel Attention with Adaptive Global Pooling (short for CAAGP).

Our proposed CAAGP consists of three main components, i.e., improved self-attention module, adaptive global pooling module and responses generation module. Firstly, the improved self-attention is responsible for spatial long-range information modeling, which provides adaptive weights for adaptive global pooling. Then the adaptive weights generated by improved self-attention is utilized by adaptive global pooling to generate the initial channel responses. Finally, the responses generation module equipped with fully connected layers are applied to generate the final

*Corresponding author.

channel-wise responses on the basis of adaptive global pooling results. In terms of global information modeling, self-attention could outperform all of other methods, however limited for its quadratic computational complexity, which could significantly increase the GPU memory and computation burdens. Inspired by the dimensionality decomposition idea in depthwise separable convolutions [4, 16], we improve and factorize self-attention [35] into $x$-direction and $y$-direction respectively, aimed to reduce the computational complexity from quadratic to linear. Our experiments have proved that long range dependencies could be modeled through the improved self-attention, which is similar to common self-attention, and the performance of backbone has been significantly improved. Moreover, the proposed CAAGP is a plug-and-play and lightweight module, which could be conveniently inserted into many function blocks, preserving more spatial and fine-grained information before generating channel-wise responses, compared with channel attention.

The main contributions are summarized as follows:

• We propose an efficient and effective method named CAAGP, a plug-and-play and lightweight module, which could simultaneously aggregate the spatial information into channel attention.

• We improve the self-attention, aggregating spatial information from $x$ and $y$ directions respectively, that effectively model spatial long-range dependencies with little calculation cost.

• Extensive experiments have been conducted to verify the effectiveness of our proposed method.

## 2. Related Work

### 2.1. Segmentation networks

For the first time, Long et al. [19] proposed FCN on the basis of CNN, completely abandoning the FC layer, which is usually used in CNNs. FCN overcomes the defects of patch-based segmentation method, reducing computational redundancy and improving global feature extraction, and truly realizes end-to-end training and pixel-to-pixel prediction. In the field of biomedical image segmentation, Ronneberge et al. [27] improved the FCN [19] and firstly proposed the encoder-decoder network(i.e., 2D U-Net), including an encoder for extracting features and an decoder for restoring resolution. 2D U-Net achieved excellent performance in biomedical image segmentation tasks, and consequently various extensions of U-Net (such as U-Net++ [39], U-Net3+ [13] and U-Net-Attention [22]) were developed for biomedical image segmentation.

### 2.2. Attention mechanisms

Many attention mechanisms have been proposed and applied to CNNs to improve the networks performance, and we roughly classified them into three categories.

**Channel attention** Hu et al. [12] firstly introduced the concept of channel attention, focused on the channel relationship and proposed a novel architectural unit, termed as the "Squeeze-and-Excitation" block [12], adaptively recalibrating channel-wise feature responses by explicitly modeling interdependencies between channels. Following the principle of channel attention, Li et al. [17] proposed a dynamic selection mechanism in CNNs that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information, in which multiple branches with different kernel sizes are fused using channel attention, yielding different sizes of the effective receptive fields of neurons in the fusion layer. Qin et al. [26] shared a similar idea with SK-Net from [17], which generates responses of different receptive fields based on channel attention.

**Naive spatial attention** Woo et al. [36] proposed Convolutional Block Attention Module (CBAM), a simple yet effective attention module, for feed-forward convolutional neural networks. The CBAM [36] module sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement. Meanwhile, Park et al. [23] proposed a simple and effective attention module, named Bottleneck Attention Module (BAM), which can be integrated with any feed-forward convolutional neural networks. The BAM [23] module infers an attention map along two separate pathways, channel and spatial. Both the CBAM [36] and BAM [23] incorporate the spatial attention, utilizing convolutional operation with large kernels, which merely could compute the local dependencies, therefore we term it as naive spatial attention.

**Self-attention** Vaswani et al. [33] proposed a new simple network architecture in NLP tasks, the Transformer, based solely on self-attention, dispensing with recurrence and convolutions entirely, and then Wang et al. [35] firstly introduced self-attention into computer vision tasks, presenting non-local operations as a generic family of building blocks for capturing long-range dependencies. As self-attention [35] has the quadratic computational complexity, which is unacceptable for GPU memory and computational resources, there are several related works [15, 7] on the application and optimization of self-attention [35]. It is a primary research point that simultaneously achieve global information modeling and reduce the complexity of the self-attention [35].

### 2.3. Preprocessing methods

Window truncation is a display technology of CT (i.e., Computed Tomography) and MRI (i.e., Magnetic Reso-

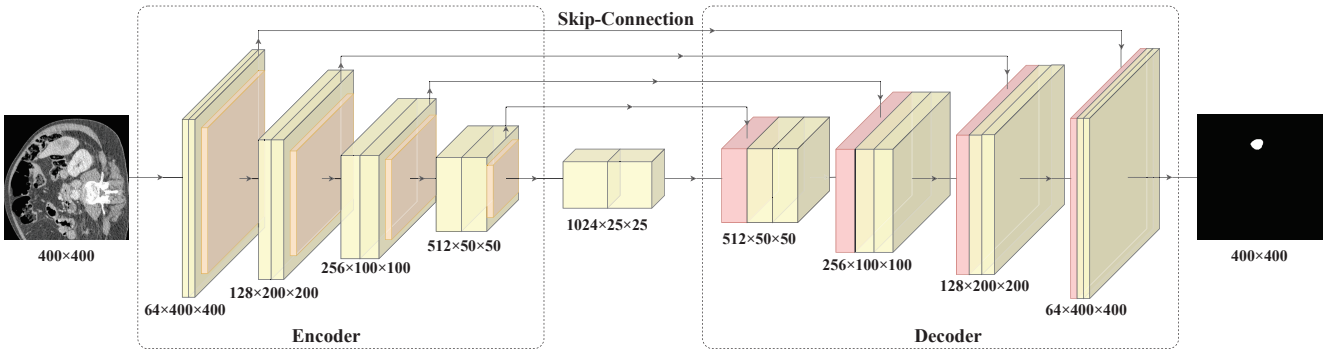


Figure 1. **U-Net architecture.** U-Net [27] is a U-shaped network, comprised of an encoder sub-network for feature extraction and an decoder for spatial resolution recovery. Feature extraction is mainly realized by cascading convolution layers, and spatial resolution restoration is realized by cascading deconvolutions.

nance Imaging) images usually used by doctors to observe normal tissues or lesions, which includes window width and window level [38]. Generally speaking, window level is the center position of Hu value of CT image target to be displayed, and window width is the display range of Hu value centered on this window level, where Hu value reflects the different degree of absorption of radiation by different tissues in CT images. Considering different tissue structures or lesions have different Hu values, it is necessary to choose the Hu reference value of the tissue as the window level, and then choose the appropriate window width to obtain the best display effect, when observing the details of a certain tissue structure.

## 3. Method

### 3.1. Preprocessing

On the basis of window truncation operation, we refer to the preprocessing method of removing mean energy in [38] as our preprocessing method, which removes average energy by statistical methods, effectively resolving the inconsistency in data. This preprocessing method could ensure the uniformity of liver gray distribution among different patients and enhance the segmentation accuracy, eventually improving the segmentation effect. Through the operation above, the gray values difference between patients could be reduced to a certain range.

### 3.2. Backbone

U-Net [27] is a U-shaped segmentation network, one of the classical networks in biomedical image segmentation tasks. The architecture of U-Net [27] is shown in Figure 1, which consists of two parts, the encoder sub-network and the decoder sub-network, hence named U-Net attributing to the U-shaped structure.

As a pioneering work in biomedical image segmentation, U-Net has been modified by many researchers to achieve further performance improvement in biomedical segmentation and other segmentation tasks. The special characteristic of U-Net is the U-shaped structure, containing encoder sub-network responsible for down-sampling and extracting features, and the decoder sub-network responsible for up-sampling and restoring image spatial information. The U-shaped network also contain the vital skip-connection to combine the low-level features from encoder and the high-level features from decoder, which could supplement the detail information in up-sampling process. Therefore, we could also utilize the U-shaped structure and improve the architecture by simply adding a residual [11, 37] path with our module, parallel to the convolutional blocks, in encoder sub-network, which is for feature extraction. In addition, naked U-Net is not sufficient to verify the effectiveness of proposed method, thus we attempt to decorate other segmentation networks, U-Net-Attention[22], U-Net++[39], U-Net3+[13], Deeplabv3+[2] and CE-Net [9], with proposed method and the experimental results are shown in section 4.4.

### 3.3. Adaptive global pooling

The global pooling is often used in channel attention [12] to encode spatial information globally, however difficult to preserve spatial fine-grained information, which is essential for liver tumor segmentation, especially for small tumors. Similarly, spatial attention utilize global pooling in channel-dimension, failing to concentrate on different features from different channels. Actually, the liver tumor segmentation has high requirements for spatial fine-grained information, which is often not paid attention to by existing models. The original design of U-Net is difficult to capture fine-grained information. Even though with the channel attention from SE-Net [12], which focuses on contribution of different channels, the calculation of global average pooling is too coarse to retain necessary spatial fine-grained information. It is necessary to incorporate contribution of channel and spatial fine-grained information simultaneously. To encourage attention blocks to capture fine-grained and long-

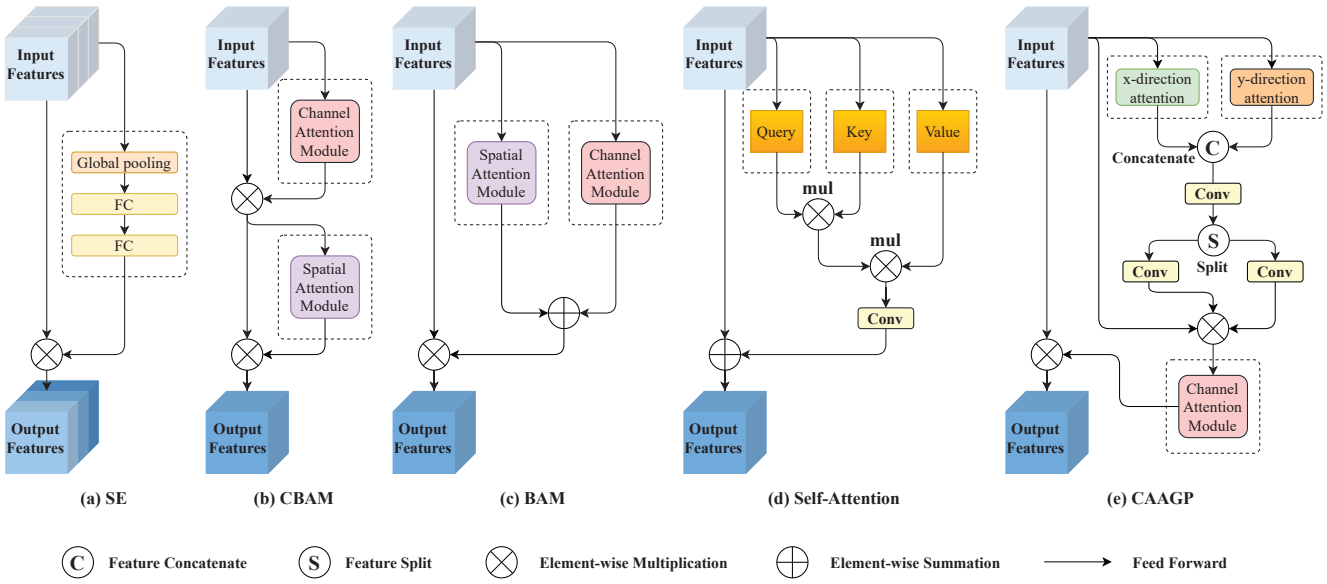


Figure 2. **Schematic comparison of some attention mechanisms and our proposed CAAGP.** (a) SE [12] is the Squeeze-and-Excitation block. (b) CBAM [36]. Convolutional Block Attention Module. (c) BAM [23]. Bottleneck Attention Module. CBAM and BAM simultaneously incorporated the channel attention and naive spatial attention. (d) Self-Attention [35], which is the first attempt to apply the self-attention in NLP to computer vision fields. (e) CAAGP. Channel Attention with Adaptive Global Pooling.

range dependencies spatially while generating the channel-wise responses, we rethink the structure of channel attention and propose the adaptive global pooling. The overview structure is shown in Figure 2(e) and the detailed structure is shown in Figure 3. Our adaptive global pooling adding global spatial information modeling before channel attention, which could preserve fine-grained information prior to global pooling, essential for small tumors segmentation.

Initially, we intend to generate spatial long-range dependencies with self-attention because of its excellent performance, while it is unacceptable for its memory costs and computational burdens because of the quadratic computational complexity. Moreover, self-attention is prone to optimization difficulties, while our improved self-attention could divide the whole optimization task into two directions, reducing the optimization pressure and computational complexity. Hence, we decide to propose an effective and efficient method to replace the self-attention [35] with little calculation cost, meanwhile capturing spatial and fine-grained information as more as possible.

Inspired by the dimensionality decomposition idea in depthwise separable convolutions [4, 16], we rethink the mechanism of self-attention, and factorize it into $x$ and $y$ directions in spatial dimension, consequently reducing the computational complexity from quadratic to linear. The decomposed self-attention can effectively capture long-range dependencies from $x$ and $y$ directions respectively, approximating the global modeling effect of self-attention through weighting the identity feature maps with these two directional attentions. We will describe it in detail in section 3.4.

Last but not least, the proposed adaptive global pooling is a plug-and-play and lightweight module, which could be conveniently inserted into many function blocks. As shown in Figure 3, we could improve the backbone by simply adding a residual [11, 37] path with our proposed CAAGP module parallel to the $k$-th layer of the encoder.

### 3.4. Improved self-attention

In the CAAGP module, the critical component is the improved self-attention, as shown in Figure 3, which effectively encodes long-range spatial dependencies along $x$-direction and $y$-direction respectively, consequently factorizing the self-attention into two linear operations and easing the memory and computational burdens.

Specifically, given the input $\mathbf{X}$=[$x_1$, $x_2$, ..., $x_c$, ...], we use two one-dimensional spatial pooling to encode global spatial information along the $x$-direction and the $y$-direction, respectively. The output of $x$-direction attention and $y$-direction attention can be formulated as:

$$z_h(c,h) = \mathbf{F}_{sq\_w}(x) = \frac{1}{W} \sum_{0 \le i \le W-1} x(c,h,i) \quad (1)$$

$$z_w(c,w) = \mathbf{F}_{sq\_h}(x) = \frac{1}{H} \sum_{0 \le j \le H-1} x(c,j,w) \quad (2)$$

where H and W represent the spatial resolution of feature maps, $c$ indicates the $c$th channel of the feature maps, and $h$ and $w$ represents the $h$th row and $w$th column respectively.

The above two equations aggregate features along the two spatial directions respectively, different from SE-Net

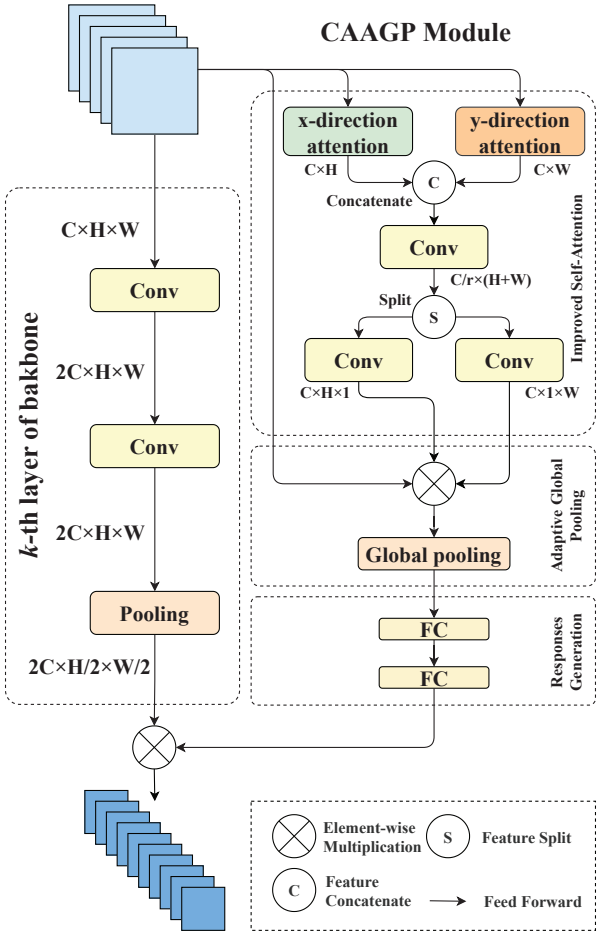


Figure 3. **The backbone decorated with CAAGP.** Our proposed CAAGP consists of three main components, i.e., improved self-attention module, adaptive global pooling module and responses generation module. We could improve the encoder of the backbone by simply adding a residual [11, 37] path with our proposed CAAGP module, which is parallel to the *k*-th layer of the encoder.

[12], which squeezes all the spatial dimension, which is coarse to retain little fine-grained information. Actually, we have simplified and decomposed the complete self-attention [35] into two linear operations through the above operations(i.e., Eq. 1 and Eq. 2). This method preserves the spatial information of the one direction while modeling the long-range dependencies information of the other direction, retaining more spatial and fine-grained information, which could be useful for small objects segmentation.

We then merge the long-range dependencies captured from both directions to model the long-range dependencies for the entire spatial dimension. Specifically, we firstly concatenate the directional attention results in Eq. 1 and Eq. 2, and then feed them into a shared $1 \times 1$ convolution.

$$\tilde{\mathbf{X}} = Conv(C[z_h(c), z_w(c)]) \tag{3}$$

where $C[\cdot, \cdot]$ denotes the concatenation operation, and $Conv(\cdot)$ denotes the $1 \times 1$ convolution operation activated by ReLU activation function [8] followed by BN (i.e., Batch Normalization) [14]. $\tilde{\mathbf{X}}$ is the generated feature maps that encode spatial long-range dependencies from the whole image.

Afterwards, we split $\tilde{\mathbf{X}}$ along the spatial dimension into two separate tensors $\tilde{\mathbf{X}}_h$ and $\tilde{\mathbf{X}}_w$. Another two operations of feature extraction $Conv_h$ and $Conv_w$, activated by Sigmoid [10] activation function, are utilized to separately generate $weight_h$ and $weight_w$

$$weight_h = Sigmoid(Conv_h(\tilde{\mathbf{X}}_h)) \tag{4}$$

$$weight_w = Sigmoid(Conv_w(\tilde{\mathbf{X}}_w)) \tag{5}$$

where the $weight_h$ and $weight_w$ are weights calculated for rows and columns respectively. The adaptive weights(i.e., $weight_h$ and $weight_w$) generated by improved self-attention will then be fed into adaptive global pooling module to model the spatial long-range dependencies, and weighted the spatial features, yielding $\mathbf{Y}$=[$y_1$, $y_2$, ..., $y_c$, ...],

$$\mathbf{Y} = \mathbf{X} \cdot weight_h \cdot weight_w \tag{6}$$

where $weight_h \in R^{C \times H \times 1}$ is utilized to weight each row in identity $\mathbf{X} \in R^{C \times H \times W}$, and $weight_w$ is for each column. The $\mathbf{Y}$ are mediate feature maps, capturing spatial long-range dependencies along both $x$ and $y$-directions, through improved self-attention and adaptive global pooling. Finally, we will feed the $\mathbf{Y}$ into responses generation module to generate channel-wise responses. The responses generation module consists of two cascaded fully connected layers activated by ReLU [8], as shown in Figure 3. Our CAAGP block not only generates channel-wise responses, but also considers captures the spatial long-range dependencies with little calculation cost. We generate attention maps along both the $x$ and $y$ directions, and aggregate them to model global spatial long-range dependencies, preserving the spatial and fine-grained information before generating channel-wise responses, which is necessary for small tumors segmentation.

As shown in Algorithm 1, we describe the process of self-attention in pseudocode. With the feature maps input X $\in R^{C \times H \times W}$, it is easy to calculate the algorithm complexity of self-attention, which is $\mathbf{O}(C \times HW \times HW + C \times HW \times HW)$. Assuming the number of channels $C$ is constant, and the final algorithm complexity of self-attention will be simplified as $\mathbf{O}((HW)^2)$. From Algorithm 2, we could easily obtain the algorithm complexity of our proposed improved self-attention, i.e., $\mathbf{O}(H + W)$, which is much less than the complexity of self-attention. Assuming $H$ is equal to $W$, improved self-attention will be optimized by two exponential levels compared to self-attention, with respect to algorithm complexity, i.e., $H^4 \rightarrow H$.

**Algorithm 1:** Implementation of self-attention

**Input:** $\mathbf{X} \in R^{C \times H \times W}$
**Output:** $\mathbf{Y} \in R^{C \times H \times W}$
• Calculate the correlation matrix **M**:
**for** *each vector* $x \in R^{C}$ *in* ***X*** **do**
  **for** $0 \leq i \leq H$ **do**
    **for** $0 \leq j \leq W$ **do**
      Calculate the correlation coefficient between x and **X**(:, $i$, $j$), by inner product, as the $i$-th, $j$-th element of correlation matrix **M**;
    **end**
  **end**
**end**
• Weight **X** with correlation matrix **M**:
**for** $0 \leq c \leq C$ **do**
  **for** *each point* $p$ *in* ***X****(c, :, :)* **do**
    Update the value of $p$, the new value is calculated through weighting the values of all points in **X**(c, :, :) by correlation coefficients from $\mathbf{M} \in R^{HW \times HW}$;
  **end**
**end**
• The updated **X** is the result **Y** to output.

There is another perspective to analyze our method. Actually, the core our algorithm is to extend the channel attention in channel dimension to two directions(i.e., $x$-direction and $y$-direction) in the spatial dimension. improved self-attention of CAAGP is operated in two directions respectively, extending the channel adaptive weighting operator to pixel level adaptive weighting, and the main difference with channel attention is that the channels is replaced with pixels. Channel attention has linear algorithm complexity, so does our proposed method. Because CAAGP inherits the characteristic of channel attention, the network could obtain the global dependencies in one direction while preserving the details in the other direction, generating the attention maps to protect fine-grained information. The effectiveness of our algorithm will be proved in section 4.

### 3.5. Loss function

**Weighted cross entropy** The image segmentation task can be considered as a pixel-level classification problem, where it is suitable to utilize cross entropy [6] to supervise and guide back propagation.

In liver tumor segmentation task, there is a serious imbalance between tumors and background, which is quite different from canonical segmentation tasks. If used ordinary cross entropy to calculate the loss, background will occupy the most, while the small liver and tumor have little contribution to the loss, which will lead to the update

**Algorithm 2:** Implementation of improved self-attention

**Input:** $\mathbf{X} \in R^{C \times H \times W}$
**Output:** $\mathbf{Y} \in R^{C \times H \times W}$
• Calculate $x$-direction attention $z_h \in R^{C \times H}$:
  Preserve the $H$ dimension;
  Pooling the $W$ dimension;
• Calculate $y$-direction attention $z_w \in R^{C \times W}$:
  Preserve the $W$ dimension;
  Pooling the $H$ dimension;
• Model the long range dependencies by combining $z_h$ and $z_w$ ;
• Generate the final attention maps, $weight_h$ and $weight_w$ in $x$ and $y$ directions;
• Weight the **X** with $weight_h$ and $weight_w$.
• The weighted **X** is the result **Y** to output.

of network parameters towards the direction of optimizing background segmentation with poor segmentation effect. In order to solve the above problem, we introduce the weights into cross entropy loss function, termed weighted cross entropy, which sets different weights for different categories, so as to emphasize the significance of this category for network classification, thus updating network parameters in the correct direction. The formula of weighted cross entropy function is:

$$CE(p, \hat{p}) = -\frac{1}{N}\sum_{i=1}^{N}[w_p \cdot plog(\hat{p}) + w_{(1-p)} \cdot (1-p)log(1-\hat{p})] \tag{7}$$

where $N$ indicates the number of pixels per slice, $w_p$ and $w_{1-p}$ indicate the weights for target and background respectively, and p and $\hat{p}$ represent the ground truth and prediction respectively. In experiments, we empirically set the weight of the foreground to $0.75$ and the weight of the background to $0.25$.

## 4. Experimental results

In this section, we conduct extensive experiments to verify the effectiveness of our proposed methods in liver tumor segmentation. In experiments, we use Python as the primary programming language and Pytorch [24] as the deep-learning framework. All models are implemented with the Pytorch framework trained and tested on GTX 1080Ti.

We utilize the LiTS2017 [38] training dataset as our train and quantitative test dataset, the 3D-IRCADB [5] and the dataset provided by Shandong Provincial Hospital as qualitative test dataset. LiTS2017 dataset is composed of CT images of 131 patients. Although there are only 131 patients, the 3D CT image is composed of thousands of 2D slices,

Table 1. **Segmentation results of several algorithms on LiTS2017 dataset.** Our method performs the best.

| Algorithm | Dice | Jaccard | Precision | Recall | F-score | Hausdorff | RVD |
|---|---|---|---|---|---|---|---|
| U-Net [27] | 0.7524 | 0.6233 | 0.8785 | 0.6858 | 0.7524 | 75.10 | -0.1512 |
| U-Net-Attention [22] | 0.7721 | 0.6444 | 0.8360 | 0.7392 | 0.7721 | 72.39 | -0.0751 |
| CE-Net [9] | 0.8102 | 0.7081 | 0.8703 | 0.7849 | 0.8102 | 57.65 | -0.0636 |
| U-Net++ [39] | 0.8317 | 0.7253 | 0.8306 | 0.8489 | 0.8317 | 86.91 | +0.0505 |
| U-Net3+ [13] | 0.8197 | 0.7228 | 0.8172 | 0.8284 | 0.8197 | 47.67 | +0.0169 |
| U-Net+CAAGP | 0.8412 | 0.7437 | 0.8393 | 0.8484 | 0.8412 | 70.32 | +0.0200 |

which is a very representative and large liver tumor segmentation dataset to the best of our knowledge. According to our statistics, among all the slices of 131 patients, there are 19,163 slices containing liver and 7,190 slices containing tumor. The total number of slices in LiTS2017 is over 30,000, the resolution of which is $512 \times 512$. At the same time, we also made statistical distribution on the resolution of liver and tumor. For all slices, the spatial resolution of liver fluctuated in the range of $0 \sim 50,000$ pixels, and that of tumor fluctuated in the range of $0 \sim 16,000$ pixels. Therefore, the dataset we used is representative in both sample diversity and scale diversity. In addition, we utilize the dataset provided by Department of Medical Imaging, Shandong Provincial Hospital as additional test dataset.

We demonstrate the effectiveness of the CAAGP on liver tumor segmentation, and have trained U-Net [27], U-Net-Attention [22], CE-Net [9], U-Net++ [39] and U-Net3+ [13] for comparisons. To evaluate algorithm performance, we adopted the evaluation metrics such as Dice [31] and Jaccard [31], which are the key evaluation metrics in biomedical image segmentation tasks, along with the Precision [25], Recall [25], F-score [28], Hausdorff(i.e., Hausdorff distance [40]) and relative volume difference(i.e., RVD [1, 34]).

### 4.1. Comparison with other segmentation methods

In this section, CAAGP was firstly inserted into U-Net, and several experiments were conducted to compare the performance with some outstanding biomedical image segmentation algorithms. Actually, our method outperformed all the compared approaches and achieved excellent results, as shown in Table 1, Figure 4, Figure 5 and Figure 6.

Our CAAGP significantly outperforms the U-Net [27] nearly 9% on Dice [31] and Jaccard [31] coefficients, which are the important metrics for evaluating the overlap ratio between images in biomedical image segmentation tasks. In addition to our proposed method, the one that achieves the best results on these two evaluation metrics is U-Net++ [39], while our method still outperformed it by two points w.r.t. the Dice [31] and Jaccard [31] metrics. Beneficial from the proposed adaptive global pooling in our method, which could preserves spatial and fine-grained information before generating channel attention, our algorithm has achieved superior performance than other algorithms in predicting the location of mask, with high overlap ratio.

In addition, we evaluated the false positive and false negative predictions of each algorithm with respect to Precision [25] and Recall [25] metrics, and our proposed CAAGP has outperformed all of other compared algorithms. It is obvious that our method is superior to other methods with respect to false positive and false negative prediction. When we utilize the F-score [28] which is the weighted summation of Precision [25] and Recall [25] to evaluate the performance, our algorithms outperform the existing algorithms significantly, even more than one point better than the optimal performance of any method other than ours. The accurate predictions could be attributed to the channel attention mechanism with the adaptive global pooling, which simultaneously aggregates spatial long-range dependencies and channel information, augmenting the representation ability of network.

Hausdorff distance [40] and RVD [34] are used to evaluate the mismatch between the prediction and ground truth. For the Hausdorff distance [40], CAAGP actually achieves the inferior performance to U-Net3+ [13], while achieving comparable performance w.r.t. RVD. In spite of the inferior performance to U-Net3+ [13] in evaluation of the mismatch between the prediction and ground truth, the CAAGP has outperformed the U-Net3+ [13] significantly according to other evaluation metrics, especially the Dice and Jaccard, which are the critical metrics in biomedical image segmentation.

Last but not least, we test our model trained on LiTS2017, on 3D-IRCADB [5] and the dataset provided by Department of Medical Imaging, Shandong Provincial Hospital, to test the generalization performance of our proposed method, and the results are shown in Figure 4. The experimental results has shown that our proposed method still achieved excellent test results on the additional test dataset which has different data distribution from LiTS2017, proving the generalization performance of our model.

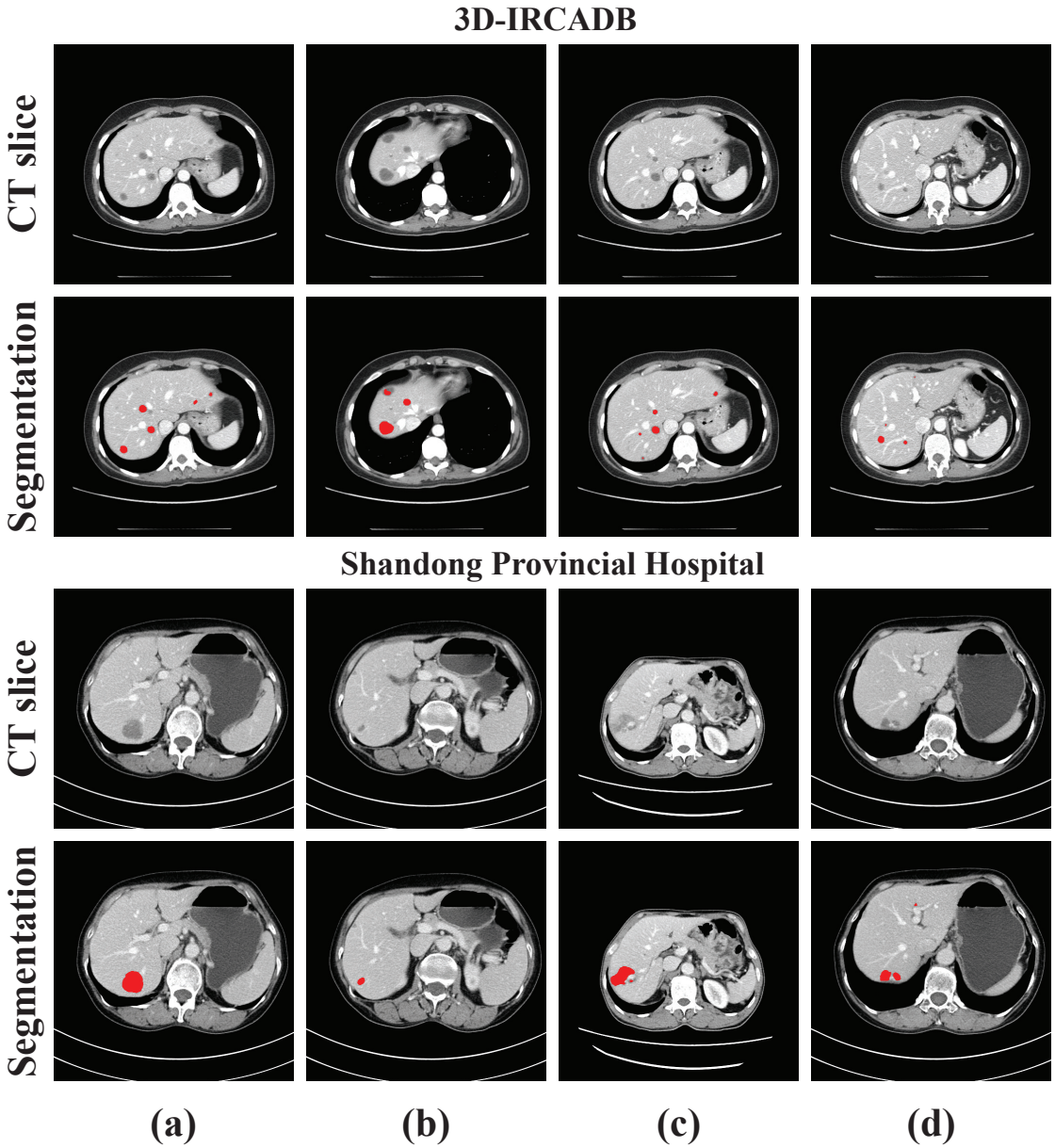


Figure 4. **Segmentation results of CAAGP on additional qualitative test dataset.** The first row exhibits the raw CT slices, and the second row exhibits the segmentation results. The red areas in the image indicate the tumors predicted by CAAGP. To prove the generalization performance of our model trained on LiTS dataset, we test our proposed CAAGP on the 3D-IRCADB [5] and the dataset provided by Department of Medical Imaging, Shandong Provincial Hospital. The proposed method has a strong ability to capture small target tumors and a high segmentation accuracy. Beneficial from the adaptive global pooling, which could preserves fine-grained information before global pooling, our proposed method has achieved excellent segmentation results on small tumors.

### 4.2. Comparison with other attention mechanisms

As shown in Table 2, we compare the segmentation results of U-Net [27] with several attention algorithms on LiTS2017 dataset. Intuitively, our proposed CAAGP performs the best on the key metrics, including Dice and Jaccard.

U-Net [27] achieves the worst performance, as it is a naked architecture without any attention mechanisms. Firstly, we added channel attention to U-Net to construct the U-Net+SE, and the results showed almost 1.5 and 3 points increase on Dice and Jaccard metrics, compared to U-Net [27]. Secondly, we added the two attention mechanisms, CBAM [36] and BAM [23], into U-Net, of which the channel attention and spatial attention are in parallel and cascade connection style respectively. The U-Net+BAM and

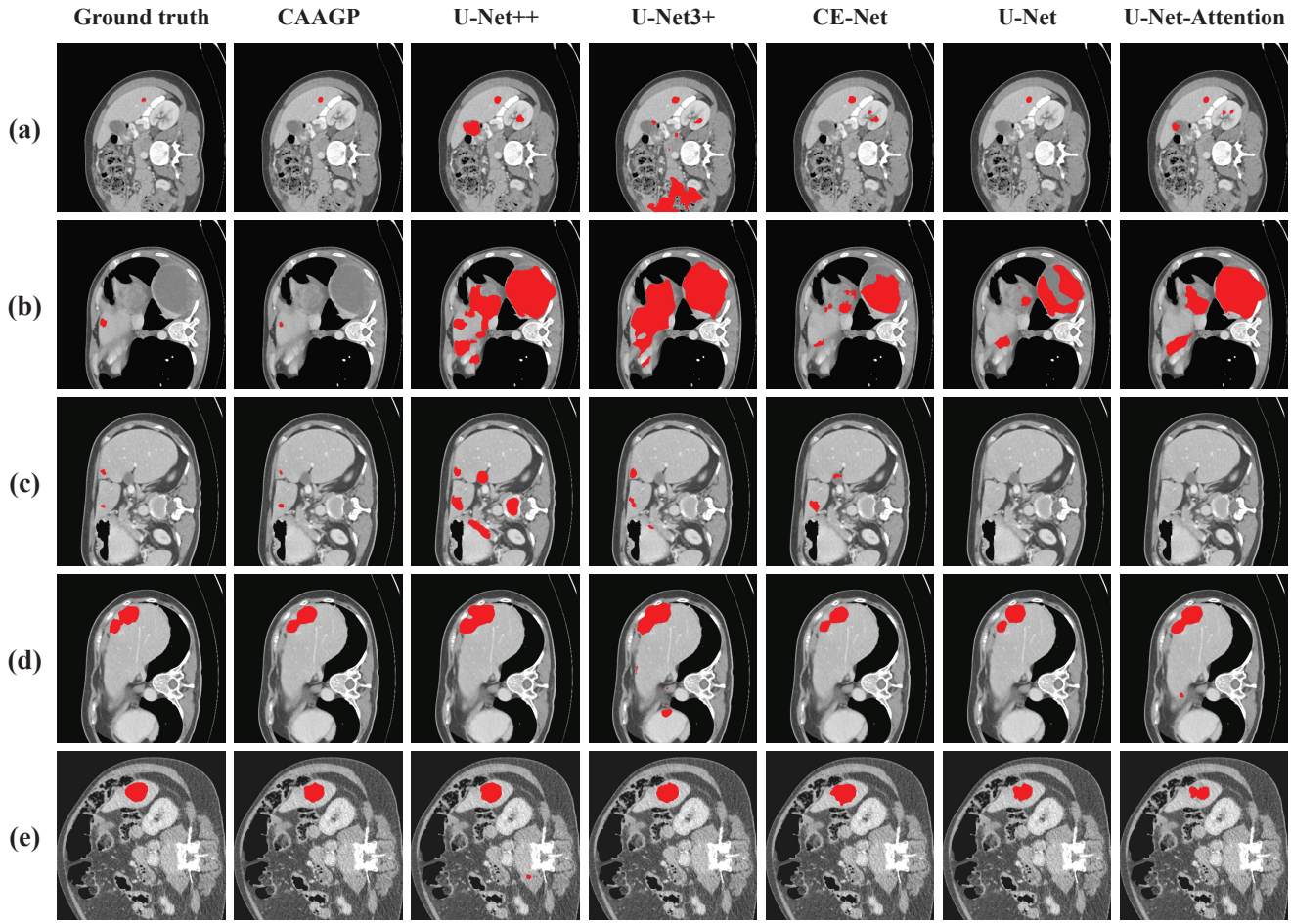


Figure 5. **Comparisons with other algorithms on the quantitative test dataset and our CAAGP has achieved the best performance.** Each row represents one sample, and each column represents Ground truth and the corresponding algorithm prediction results. With respect to (a), (b) and (c), which are the samples containing very small tumors, without obvious features, other segmentation algorithms either generated large areas of false positive predictions or failed to predict the small tumors, while the proposed CAAGP accurately located and segmented the tumors despite the extremely small size. Especially for sample (b), all of the algorithms other than our proposed CAAGP have mispredicted a large number of pixels. With respect to (d) and (e), the samples which have larger tumors, our proposed CAAGP still has outperformed all of other compared algorithms significantly, either in terms of the predicted number of tumors, or the locations of tumors. The predictions of CAAGP shown above possessed better segmentation edges than the predictions of other algorithms.

U-Net+CBAM both achieved 5 points increase in terms of both Dice and Jaccard metrics, compared to U-Net [27], whether it is parallel or cascade connection style. However, neither CBAM or BAM considers to aggregate the channel attention and spatial attention with an effective way, simultaneously taking channel information and spatial information into account, and they are in the naive parallel or cascade connection style respectively. To demonstrate the effectiveness of our idea, which incorporates the spatial information into the channel attention, by means of adaptive global pooling, we added the naive spatial attention based on the U-Net-SE, constructing the U-Net+SE+spatial. There is no obvious point of increase as shown in the test results of U-Net+SE+spatial, compared to U-Net+CBAM and U-Net+BAM, which is only 0.5 and 2 points w.r.t. Dice and Jaccard metrics respectively. We supposes that this resulted from the local computing of naive spatial attention. Hence, we conducted the experiment of our proposed U-Net+CAAGP, which encodes the spatial long-range dependencies globally, and incorporates them into channel-wise responses generation. Our proposed method has outperformed all of the attention mechanisms, and achieved almost 2 and 4 points increase in terms of Dice and Jaccard metrics respectively, compared to U-Net+BAM and U-Net+CBAM.

Table 2. **Segmentation results of several attention algorithms on LiTS2017 dataset.** Our method performs the best on the key metrics. Best performance is bold-faced.

| Algorithm | Dice | Jaccard |
|---|---|---|
| Non-Attention | 0.7524 | 0.6233 |
| SE [12] | 0.7667 | 0.6566 |
| BAM [23] | 0.8158 | 0.7077 |
| CBAM [36] | 0.8178 | 0.7080 |
| SE(+spatial) | 0.8207 | 0.7225 |
| **CAAGP** | **0.8412** | **0.7437** |

Table 3. **Segmentation results under different settings of the proposed CAAGP on LiTS2017 dataset.** As we take both $x$-direction and $y$-direction attention into account (CAAGP), our approach performed the best.

| Algorithm | Dice | Jaccard |
|---|---|---|
| Non-Attention | 0.7524 | 0.6233 |
| CAAGP($x$-direction) | 0.8264 | 0.7245 |
| CAAGP($y$-direction) | 0.8246 | 0.7265 |
| **CAAGP(both-directions)** | **0.8412** | **0.7437** |

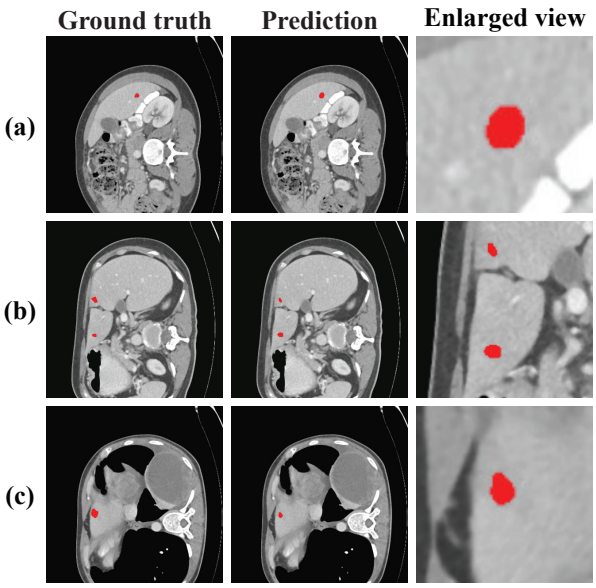


Figure 6. **Prediction examples of CAAGP on LiTS2017 dataset.** Each row represents a sample, and the three column represent Ground truth, the predictions of our proposed CAAGP and Enlarged view of the predictions respectively. Our CAAGP has achieved excellent performance on small objects. Our proposed method could accurately locate and segment even some tumors, which are very small, without obvious features.

### 4.3. Ablation studies on different settings

To demonstrate the effectiveness and rationality of the idea of factorizing self-attention [35] into two linear attention operations along two directions respectively, we perform a series of ablation experiments, and the corresponding results of which are shown in Table 3.

Firstly, we only added global spatial long-range dependencies modeling from either $x$ or $y$ direction, constructing the U-Net+CAAGP(x-direction) and U-Net+CAAGP(y-direction) respectively, and the two also achieved almost equivalent performance, far superior to U-Net [27]. However, when both the $x$-direction and the $y$-direction attention are incorporated, CAAGP achieved the best result as bold-faced in Table 3. As shown in Figure 7, we compare the results of ablation experiments. It is obvious that test results of CAAGP are most closest to the ground truth, and there is no prediction of false negatives and false positives. Either CAAGP-x or U-Net produces more or less false positive predictions.

Liver tumors have no obvious features, such as shape, number, etc., and the types of liver tumors are also diverse, that is, tumor characteristics of different patients will be completely different, which is a great challenge for semantic segmentation. However, with the benefit of adaptive global pooling, which could preserve spatial and fine-grained information when generating channel attention, our proposed CAAGP has achieved excellent results on liver tumor segmentation tasks, especially for small tumors.

### 4.4. Studies on different bases

To further demonstrate the effectiveness and rationality of our proposed method, we incorporate CAAGP into different bases, including [22, 39, 13, 2, 9], four biomedical segmentation networks and one semantic segmentation network. A series of experiments have been conducted and the corresponding results of which are shown in Table 4.

We mainly measure the improvement of the proposed CAAGP on the original network performance from three perspectives, overlap ratio, false positive and false negative, and matching distance between the prediction and ground truth, and three metrics are applied, i.e., Jaccard, F-score and Hausdorff distance. It is obvious that there has been segmentation performance improvement in each segmentation network w.r.t. almost each metric. Although the performance of each network is improved with CAAGP, the degree of improvement is different among these networks. We speculate that the performance improvement is related to the structure of the network itself, and the naive combination of the original mechanisms with our proposed method may not contribute to the optimal performance. In the same way, CAAGP is a better combination style of channel atten-

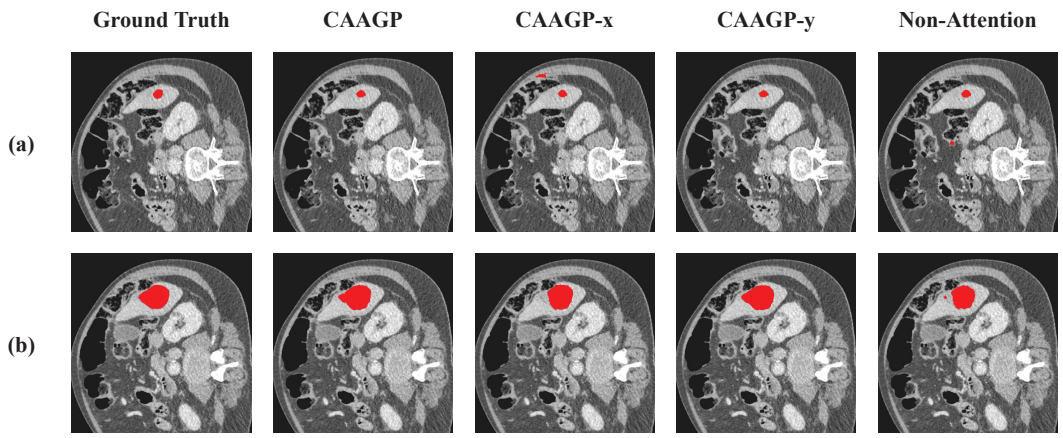


Figure 7. **Results of ablation experiments.** Test results of CAAGP are most closest to the ground truth, and there is no prediction of false negatives and false positives.

Table 4. **Segmentation results on different bases with the proposed CAAGP on LiTS2017 dataset. Not only the biomedical segmentation networks can be improved with CAAGP pattern, but also the semantic segmentation network for natural images segmentation.**

| Algorithm | Jaccard | F-score | Hausdorff |
|---|---|---|---|
| Deeplabv3+ [2] | 0.5123 | 0.6578 | 109.61 |
| +CAAGP | 0.5544 | 0.6929 | 101.40 |
| U-Net-Attention [22] | 0.6444 | 0.7721 | 72.39 |
| +CAAGP | 0.7157 | 0.8171 | 51.24 |
| CE-Net [9] | 0.7081 | 0.8102 | 57.65 |
| +CAAGP | 0.7431 | 0.8312 | 35.91 |
| U-Net++ [39] | 0.7253 | 0.8317 | 86.91 |
| +CAAGP | 0.7705 | 0.8588 | 34.85 |
| U-Net3+ [13] | 0.7228 | 0.8197 | 47.67 |
| +CAAGP | 0.7384 | 0.8349 | 36.13 |

tion and spatial attention, than naive concatenation or summation.

In addition to the biomedical image segmentation network, we also try to apply semantic segmentation network for natural images to our experiments to verify the effectiveness of our proposed method. Compared with natural images, biomedical images are relatively simple, generally only contain a single channel, and are highly correlated with each other. However, the data of biomedical image is simple while the contrast between various tissues and organs in biomedical images is low, which greatly boosts the difficulty of feature extraction. Deeplabv3+ [2] lacks the rich skip-connection structure like encoder-decoder networks(e.g., U-Net [27]), contributing to the significant block of detailed information propagation, therefore its segmentation performance has a gap compared with other biomedical image segmentation networks. Fortunately, benefiting from the strong feature extraction ability of our CAAGP, there is still a considerable performance improvement in Deeplabv3+, when decorated with CAAGP.

Through the above experiments, it is sufficient to prove the effectiveness of our proposed CAAGP in biomedical segmentation task. Meanwhile, transfer learning [32] is utilized to accelerate the training on the original network basis combined with CAAGP and further improve the performance.

## 5. Conclusion

In this paper, we propose a novel attention mechanism termed CAAGP, which could preserve the spatial and fine-grained information with the adaptive global pooling, while generating channel attention. Benefits form the adaptive global pooling, CAAGP has achieved excellent results on liver tumor segmentation, especially for small tumors. Meanwhile, we incorporate the improved self-attention into CAAGP, factorizing self-attention into $x$-direction and $y$-direction, which could effectively absorb spatial information and effectively reduce computational complexity from quadratic to linear. Extensive experiments on LiTS2017, 3D-IRCADB and the additional test datasets provided by Shandong Provincial Hospital have significantly proved the effectiveness and generalization performance of our CAAGP in liver tumor segmentation task.

## Acknowledgement

We thank the Department of Medical Imaging, Shandong Provincial Hospital for providing the additional qualitative test dataset of liver tumors.

## References


[1] Sang Hee Ahn, Adam Unjin Yeo, Kwang Hyeon Kim, Chankyu Kim, Youngmoon Goh, Shinhaeng Cho, Se Byeong Lee, Young Kyung Lim, Haksoo Kim, Dongho Shin, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiation Oncology*, 14(1):1–13, 2019. 7

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3, 10, 11

[3] Yilong Chen, Kai Wang, Xiangyun Liao, Yinling Qian, Qiong Wang, Zhiyong Yuan, and Pheng-Ann Heng. Channel-unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation. *Frontiers in genetics*, 10:1110, 2019. 1

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2, 4

[5] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016. 6, 7, 8

[6] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. 6

[7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 1, 2

[8] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 5

[9] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019. 3, 7, 10, 11

[10] Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International Workshop on Artificial Neural Networks*, pages 195–201. Springer, 1995. 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 5

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 4, 5, 10

[13] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 1, 2, 3, 7, 10, 11

[14] Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 791–800, 2018. 5

[15] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S Huang. Ccnet: Criss-cross attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2

[16] Dac-Nhuong Le, Velmurugan Subbiah Parvathy, Deepak Gupta, Ashish Khanna, Joel JPC Rodrigues, and K Shankar. Iot enabled depthwise separable convolution neural network with deep support vector machine for covid-19 diagnosis and classification. *International Journal of Machine Learning and Cybernetics*, pages 1–14, 2021. 2, 4

[17] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019. 2

[18] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. 2013. 1

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[20] Lei Mou, Yitian Zhao, Li Chen, Jun Cheng, Zaiwang Gu, Huaying Hao, Hong Qi, Yalin Zheng, Alejandro Frangi, and Jiang Liu. Cs-net: channel and spatial attention network for curvilinear structure segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–730. Springer, 2019. 1

[21] Jiajia Ni, Jianhuang Wu, Haoyu Wang, Jing Tong, Zhengming Chen, Kelvin KL Wong, and Derek Abbott. Global channel attention networks for intracranial vessel segmentation. *Computers in biology and medicine*, 118:103639, 2020. 1

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1, 2, 3, 7, 10, 11

[23] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 1, 2, 4, 8, 10

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[25] David MW Powers. Visualization of tradeoff in evaluation: from precision-recall & pn to lift, roc & bird. *arXiv preprint arXiv:1505.00401*, 2015. 7

[26] Yao Qin, Konstantinos Kamnitsas, Siddharth Ancha, Jay Nanavati, Garrison Cottrell, Antonio Criminisi, and Aditya Nori. Autofocus layer for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2018. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 3, 7, 8, 9, 10, 11

[28] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006. 7

[29] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*, 2021. 1

[30] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020. 1

[31] Vikas Thada and Vivek Jaglan. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2(4):202–205, 2013. 7

[32] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010. 11

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 2

[34] Eugene Vorontsov, An Tang, Chris Pal, and Samuel Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1332–1335. IEEE, 2018. 7

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4, 5, 10

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 4, 8, 10

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3, 4, 5

[38] Chi Zhang, Qianqian Hua, Yingying Chu, and Pengwei Wang. Liver tumor segmentation using 2.5 d uv-net with multi-scale convolution. *Computers in Biology and Medicine*, 133:104424, 2021. 3, 6

[39] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 1, 2, 3, 7, 10, 11

[40] Zhi-Qiang Zhou and Bo Wang. A modified hausdorff distance using edge gradient for robust object matching. In *2009 International Conference on Image Analysis and Signal Processing*, pages 250–254. IEEE, 2009. 7