



**HAL**  
open science

## CA-SegNet: A channel-attention encoder-decoder network for histopathological image segmentation

Feng He, Weibo Wang, Lijuan Ren, Yixuan Zhao, Zhengjun Liu, Yue-Min Zhu

► **To cite this version:**

Feng He, Weibo Wang, Lijuan Ren, Yixuan Zhao, Zhengjun Liu, et al.. CA-SegNet: A channel-attention encoder-decoder network for histopathological image segmentation. *Biomedical Signal Processing and Control*, 2024, 96, pp.106590. 10.1016/j.bspc.2024.106590 . hal-04852251

**HAL Id: hal-04852251**

**<https://hal.science/hal-04852251v1>**

Submitted on 20 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Highlights

### **CA-SegNet: A channel-attention encoder-decoder network for histopathological image segmentation**

Feng He, Weibo Wang, Lijuan Ren, Yixuan Zhao, Zhengjun Liu, Yuemin Zhu

- A novel deep learning-based CA-SegNet model for histopathological image segmentation.
- A channel-attention feature fusion module (CAFFM) that significantly improves shallow feature reuse.
- A bottleneck-structured decoder developed for better feature integration.
- Outstanding segmentation performance on both large and small-scale datasets.

# CA-SegNet: A channel-attention encoder-decoder network for histopathological image segmentation

Feng He<sup>a,b,f</sup>, Weibo Wang<sup>a,b,\*</sup>, Lijuan Ren<sup>c</sup>, Yixuan Zhao<sup>d</sup>, Zhengjun Liu<sup>e</sup> and Yuemin Zhu<sup>f,\*</sup>

<sup>a</sup>Institute of Ultra-precision Optoelectronic Instrument Engineering, Harbin Institute of Technology, Harbin, 150001, China

<sup>b</sup>Key Lab of Ultra-precision Intelligent Instrumentation, Harbin Institute of Technology, Harbin, 150001, China

<sup>c</sup>School of Software Engineering, Chengdu University of Information Technology, Chengdu, 610225, China

<sup>d</sup>Research Center of Advanced Microscopy and Instrumentation, Harbin Institute of Technology, Harbin, 150001, China

<sup>e</sup>School of Physics, Harbin Institute of Technology, Harbin, 150001, China

<sup>f</sup>CREATIS, INSA Lyon, CNRS UMR 5220, INSERM U1294, Universit de Lyon, Villeurbanne, 69621, France

## ARTICLE INFO

### Keywords:

Deep learning  
Channel attention  
Encoder-decoder  
Medical image segmentation  
Histopathological images

## ABSTRACT

Histopathological image segmentation based on encoder-decoder architectures has emerged as a pivotal research area in medical image analysis. However, due to the irrelevant information within multi-channel representations from the encoder, the coarse reuse of shallow features in skip connections may burden the learning and even adversely affect the decoder. While various variants have been developed to cope with this issue, the performance remains unsatisfactory. In this work, we propose a novel encoder-decoder architecture named CA-SegNet to address the above issue more effectively and achieve advanced histopathological image segmentation. Our novelty is twofold: firstly, a bottleneck-structured decoder is developed to improve the integration of multi-channel feature representations, and secondly, a sequence of channel-attention feature fusion modules (CAFFMs) are developed to adaptively guide the reuse of fine-grained shallow features in skip connections while learning the channel-wise dependencies. Experimental results on different publicly available histopathological image datasets demonstrate that our CA-SegNet outperforms existing state-of-the-art methods on both large and small-scale datasets.

## 1. Introduction


Histopathological image analysis is of great importance and the gold standard to assist medical professionals in diagnosing many diseases (e.g., breast cancer [1–3], colon cancer [4–6], and lung cancer [7–9]) and monitoring treatment progress. It encompasses a wide range of tasks and needs, where histopathological image segmentation is significant since it presents details such as the shape, volume, and location of lesion regions, which reveal necessary diagnostic indicators to support clinical decisions. However, the mainstream of this segmentation task lies in the manual effort of well-trained experts, which requires plenty of workloads and suffers observer confusion caused by interclass similarities and intraclass differences. Thus, automatic segmentation methods that establish computer-aided diagnosis systems and produce accurate and reliable segmentation results have been widely explored, particularly the methods based on deep learning [10–14].

Deep learning methods, especially convolutional neural networks (CNNs), have made remarkable progress and have been experimentally demonstrated to maintain state-of-the-art performance in computer vision tasks [15–19] (including medical image analysis) since the incredible success made

by the VGG model [20]. Specifically, in image semantic segmentation, it is well known that the fully convolutional network (FCN) [21] emerged as a pioneering solution for end-to-end image segmentation, soundly outperforming existing methods of the year. The torch of innovation was carried forward by SegNet [22], which features a basic encoder-decoder architecture and proposes the ingenious idea of max-pooling indices-based up-sampling to save more boundary features. The encoder with deep convolutional layers meticulously extracts intricate spatial features of targets and acquires local and global feature representations containing semantic information through the increasing receptive field produced by down-sampling. The complementary decoder integrates the feature representations output from the encoder and constructs the segmentation results (i.e., pixel-level masks), where the feature resolution is recovered via several max-pooling indices-based up-sampling.

Notably, the above-mentioned networks are developed for non-medical image segmentation and have an inevitable flaw, i.e., requiring large-scale datasets, which poses a formidable challenge for medical usage since the latter mostly involves small-scale datasets. Despite this flaw, the field of medical image segmentation has flourished mainly due to the remarkable potential unlocked by U-Net [23], which constructs the standard encoder-decoder architecture and proposes the groundbreaking and ingenious concept of skip connections between the encoder and decoder to reuse the fine-grained information in shallow features. U-Net is customarily considered the de facto standard in medical image segmentation due to its ability to achieve impressive segmentation results with relatively small-scale datasets.

\*Corresponding author

 fenghe@hit.edu.cn (F. He); wwbhit@hit.edu.cn (W. Wang);

renlijuan@cuit.edu.cn (L. Ren); zhaoyixuan@hit.edu.cn (Y. Zhao);

zjliu@hit.edu.cn (Z. Liu); yue-min.zhu@creatis.insa-lyon.fr (Y. Zhu)

ORCID(s):

Although the skip connections of U-Net preserve fine-grained spatial information contributing to higher segmentation accuracy, the crude concatenation of shallow and deep feature representations may diminish the positive effect of these skip connections on network performance. This is because an intrinsic expression for the encoder to extract target features involves the separation of the foreground from the background through multi-channel feature representations and then progressively attenuating the background activation. Thus, the features obtained from shallow layers contain irrelevant information (e.g., background noise) within multiple channels, and the irrelevant information increases with the forward of the decoder (since the features reused in the skip connection become shallower as the decoder forwards, and these features keep stacking up). This property burdens the learning of the decoder and even produces side effects to segmentation results, thereby decreasing the expected efficacy of skip connections. Moreover, the fine-grained feature representations of different channels focus differently on target features, which yields channel-wise dependencies between these representations. The channel-wise dependencies are essential in refining the integration ability of the decoder, which is not considered in standard encoder-decoder networks. While numerous efforts [24–26] have been devoted to addressing the above issues, histopathological image segmentation performance remains a large room for improvement.

In this work, we explore a refined and effective attention mechanism to avoid the detrimental impact caused by irrelevant information within multi-channel representations of shallow features in skip connections. Specifically, we propose a channel-attention encoder-decoder model (CA-SegNet) for histopathological image segmentation. The network encoder is built upon VGG16, which enables leveraging ImageNet-trained parameters based on transfer learning to avoid network overfitting. The decoder is designed with a bottleneck structure having a better ability to integrate feature representations. A sequence of novel channel-attention feature fusion modules (CAFFMs) that consider the channel-wise dependencies and eliminate irrelevant information are developed in skip connections. We validate the performance of our CA-SegNet on two different histopathological image segmentation tasks, i.e., breast cancer and colon cancer.

The main contributions of this paper are summarized as follows:

- We propose a novel encoder-decoder architecture called CA-SegNet incorporated with an elaborated attention mechanism to achieve accurate histopathological image segmentation.
- We develop a CAFFM to guide the skip connection between the encoder and decoder, which discredits undesired information in multi-channel representations of the fine-grained shallow features. It utilizes a weighted average pooling (WAP) method to subtly compute channel scores according to the weights of feature activation in each channel while learning

the channel-wise dependencies through a convolutional scaling operation that produces refined channel-attention coefficients.

- We design a bottleneck-structured decoder to integrate multi-channel feature representations from the encoder, where each convolutional unit constitutes a bottleneck structure. It is more effective than the standard decoder structure.
- We demonstrate the effectiveness of our CA-SegNet and its superior performance compared to existing state-of-the-art segmentation networks on two different histopathological image datasets with large and small scales.

The rest of this article is organized as follows. Section 2 gives a review of the related work. Section 3 describes the proposed CA-SegNet in detail. Section 4 provides experiments and results. Section 5 gives relevant discussion. Finally, Section 6 gives the conclusion of this paper.

## 2. Related work

This section first briefly reviews many typical segmentation networks developed for medical image analysis and then focuses on some relevant works attempting to leverage attention mechanisms to perfect skip connections between the encoder and decoder so as to improve the segmentation performance.

### 2.1. Medical image segmentation

Many efforts [27–30] based on deep learning have been made to achieve outstanding performance in medical image segmentation. The variants of U-Net were the most popular and successful outcomes. For instance, UNet++ [31] introduced a series of nested and dense skip connections in the encoder-decoder network to reduce the semantic gap between shallow and deep features and the deep supervision to facilitate better convergence during training. R2U-Net [32] leveraged the strengths of U-Net, residual networks, and recurrent CNNs to obtain better feature representation. MultiResUNet [33], with the encoder and decoder layers replaced by inception-like blocks with residual connections, introduced a chain of convolutional layers with residual connections into skip connections to alleviate the disparity between the encoder-decoder features. DoubleU-Net [34] combined two U-Net architectures stacked on top of each other and adopted atrous spatial pyramid pooling (ASPP) to improve its performance on various segmentation tasks. SMU-Net [35] adopted U-Net as a main network and incorporated it with an additional middle stream and an auxiliary network to learn foreground-salient and background-salient representations under the guidance of saliency maps to rich textural information for breast lesion segmentation in ultrasound Images. KiU-Net [36] parallelly connected an overcomplete convolutional network, which projects input images into a higher dimension, with U-Net through a series

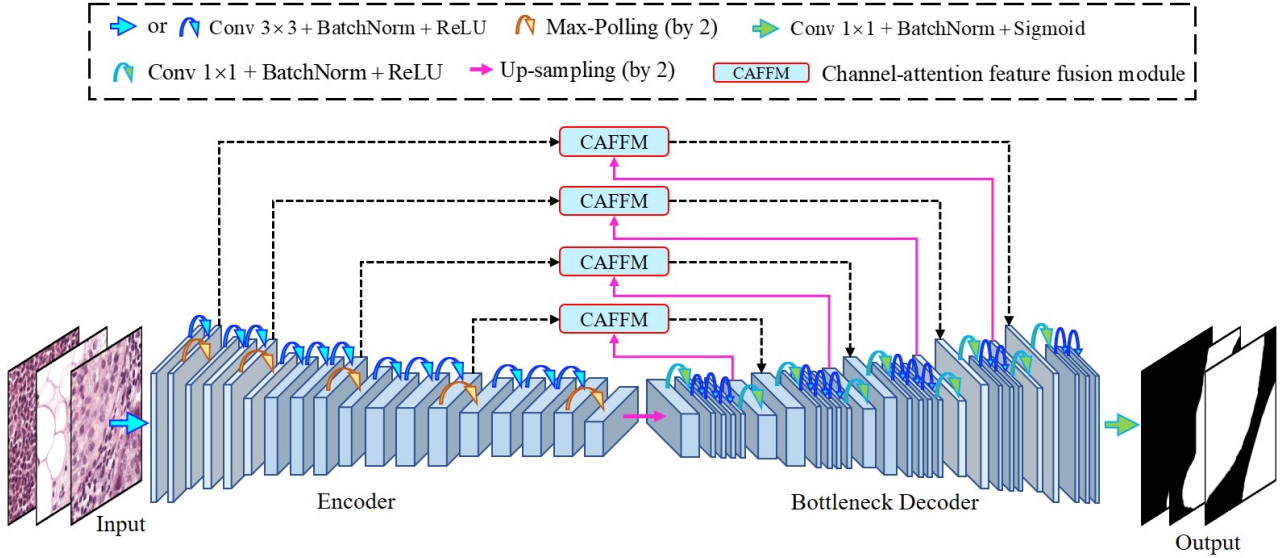


Figure 1: Architecture details of the proposed CA-SegNet.

of novel cross-residual feature blocks to identify smaller structures and segment boundary regions precisely.

While these methods have made impressive progress in medical image segmentation, they either mainly focus on reducing the semantic gap of feature maps between the encoder and decoder or improving the network feature representation to increase the network generalization and the response to tiny structures. The adverse impact of background noise-relevant information in shallow features on the segmentation performance has not been seriously considered.

## 2.2. Medical image segmentation based on attention mechanisms

Due to the properties of emphasizing regions of interest (ROIs) and filtering irrelevant features, attention mechanisms have brought considerable gains in the performance of deep learning models in medical image segmentation. In particular, researchers are keen to incorporate attention mechanisms into skip connections of encoder-decoder architectures to optimize feature fusion between shallow and deep layers. Attention U-Net [37] proposed a spatial-wise attention gate for skip connections to encourage the network to focus on target features of varying shapes and sizes while suppressing the activation of irrelevant regions. Multi-Res-Attention UNet [38] developed a hybrid skip connection-based architecture with the optimal placement of respaths, attention gates, and the usage of multi-res blocks to reduce the semantic gap between feature maps of shallow and deep layers. R2AU-Net [39] switched the basic convolutional unit of the general U-Net into a recurrent residual convolutional unit and introduced a series of attention gates into skip connections. AACA-MLA-D-UNet [40] proposed an adaptive atrous channel attention module in skip connection to sort the importance of each feature channel for retinal vessel

segmentation. CFU-Net [41] embedded two U-Nets, consisting of a partly shared encoding path and paired coarse-fine decoding paths, and designed a multilevel attention module (MLAM) that executes the multilevel information interaction to refine the feature propagation in skip connection.

Despite the fact that attention mechanisms have become increasingly popular for medical image segmentation, works on enhancing the performance of deep learning networks for histopathological image segmentation based on attention mechanisms remain scarce. Moreover, few attention mechanisms have focused on effectively discrediting the irrelevant information within multi-channel representations of shallow features in skip connections and considering channel-wise dependencies.

## 3. Methodology

Fig. 1 shows the overall architecture of the proposed CA-SegNet. Our CA-SegNet consists of a VGG16-based encoder, a bottleneck decoder, and a sequence of CAFFMs. Given the input of medical image  $I \in \mathbb{R}^{3 \times H \times W}$  with the spatial size of  $H \times W$ , the encoder gradually extracts the target features in a down-sampling manner, where the features are halved in spatial resolution and doubled in channels after each down-sampling layer to expand the receptive field and enrich feature combinations and representations. The resulting multi-scale feature maps  $F_e \in \mathbb{R}^{(H/2^e \times W/2^e) \times C}$  with channels  $C$  ( $C \in \{64, 128, 256, 512, 512\}$ ) of encoder layer  $e$  ( $e \in \{0, 1, 2, 3, 4\}$ ) are stacked into a feature pyramid  $F = \{F_0, F_1, F_2, F_3, F_4\}$  (as shown in the encoder pipeline of Fig. 1). The bottleneck decoder takes the highest feature maps  $F_4$  of the feature pyramid as the first input and gradually integrates multi-channel feature representations to construct the segmentation result (i.e., pixel-level mask with target regions annotated as 1 and background as 0) in an up-sampling manner, where the remaining multi-scale feature

maps of the feature pyramid are correspondingly input to the decoder after each up-sampling operation with the guide of a sequence of CAFFMs.

### 3.1. VGG16-based encoder

We adopt all five convolution units (see more details in Table 1) of VGG16 as our encoder. The motivation mainly falls into two benefits: 1) leverage the ImageNet-trained parameters (publicly available) to initialize the encoder so as to avoid small dataset overfitting and accelerate model convergence; 2) leverage its well-designed target feature extraction structure (since a good enough feature extraction structure is necessary to achieve accurate classification performance of VGG16). Inspired by SegNet [22], the inverse max-pooling (i.e., max-pooling indices-based up-sampling) is used to conduct the up-sampling operation in our CA-SegNet, which experimentally achieved better performance than the interpolation up-sampling and has much lower computational complexity than the transposed convolution (which introduces additional parameters requiring training). Thus, the max-pooling indices (i.e., the locations of the maximum feature value in each pooling window) of each down-sampling process of the encoder are saved and later used in the corresponding up-sampling process in the decoder.

**Table 1**  
VGG16-based encoder

Convolution Unit	Structure
0	[Conv $3 \times 3$ + BN + ReLU, C = 64] $\times 2$
1	Maxpool $2 \times 2$ [Conv $3 \times 3$ + BN + ReLU, C = 128] $\times 2$
2	Maxpool $2 \times 2$ [Conv $3 \times 3$ + BN + ReLU, C = 256] $\times 3$
3	Maxpool $2 \times 2$ [Conv $3 \times 3$ + BN + ReLU, C = 512] $\times 3$
4	Maxpool $2 \times 2$ [Conv $3 \times 3$ + BN + ReLU, C = 512] $\times 3$ Maxpool $2 \times 2$

### 3.2. Bottleneck decoder

Different from standard encoder-decoder networks, such as U-Net [23], SegNet [22], and MultiResUNet [33], where the decoder is constructed symmetrically to the encoder, we develop a bottleneck-structured decoder (as shown in the bottleneck decoder pipeline of Fig. 1) with reduced parameters that account for computational complexity and increased network depth (and nonlinearity) that enhances the feature representation.

The more detailed structure of the bottleneck decoder is given in Table 2. The decoder contains five bottleneck units, each of which up-samples its input to twice the original resolution through ‘‘Maxunpool  $2 \times 2$ ’’ (i.e., inverse max-pooling) with a stride of 2 at the beginning. The inverse

**Table 2**  
Bottleneck decoder

Bottleneck Unit	Structure
0	Maxunpool $2 \times 2$ Conv $1 \times 1$ + BN + ReLU, C = 64 [Conv $3 \times 3$ + BN + ReLU, C = 64] $\times 3$ Conv $1 \times 1$ + BN + ReLU, C = 512
1	Maxunpool $2 \times 2$ Conv $1 \times 1$ + BN + ReLU, C = 64 [Conv $3 \times 3$ + BN + ReLU, C = 64] $\times 3$ Conv $1 \times 1$ + BN + ReLU, C = 256
2	Maxunpool $2 \times 2$ Conv $1 \times 1$ + BN + ReLU, C = 64 [Conv $3 \times 3$ + BN + ReLU, C = 64] $\times 3$ Conv $1 \times 1$ + BN + ReLU, C = 128
3	Maxunpool $2 \times 2$ Conv $1 \times 1$ + BN + ReLU, C = 32 [Conv $3 \times 3$ + BN + ReLU, C = 32] $\times 2$ Conv $1 \times 1$ + BN + ReLU, C = 64
4	Maxunpool $2 \times 2$ Conv $1 \times 1$ + BN + ReLU, C = 32 [Conv $3 \times 3$ + BN + ReLU, C = 32] $\times 2$ Conv $1 \times 1$ , C = 1

max-pooling is achieved by mapping the pixels of the low-resolution feature map to the saved positions from the encoder, which produces a sparse high-resolution feature map, and it recovers more details of the target information (such as boundaries) due to the saved position information. The first three bottleneck units then reduce the channels of up-sampled features to 64 via a  $1 \times 1$  convolution operation, followed by a  $3 \times 3$  convolution operation repeated three times to integrate the complex features of different channels and transform the sparse up-sampled feature maps into dense maps. This  $1 \times 1$  convolution operation constructs the bottleneck structure, which increases the information interaction between different channels and reduces the parameters required for subsequent computations. As for the last two bottleneck units, the feature channels after the first  $1 \times 1$  convolution operation are reduced to 32, and the  $3 \times 3$  convolution operation is repeated twice. To output the feature maps having the dimension corresponding to the saved max-pooling indices that decide the up-sampling operation, another  $1 \times 1$  convolution operation is added at the end of each bottleneck unit to increase the feature channels, except for the last unit where the  $1 \times 1$  convolution operation with a single output channel is used to construct the pixel-level mask. All the above-mentioned convolution operations (except the last layer) are followed by batch normalization to reduce the internal covariate shift and a ReLU activation function to introduce nonlinearity.

### 3.3. Channel-attention feature fusion module

Since the features produced by the encoder are rather shallower in feature extraction level and less global in contextual semantic information as they are computed earlier

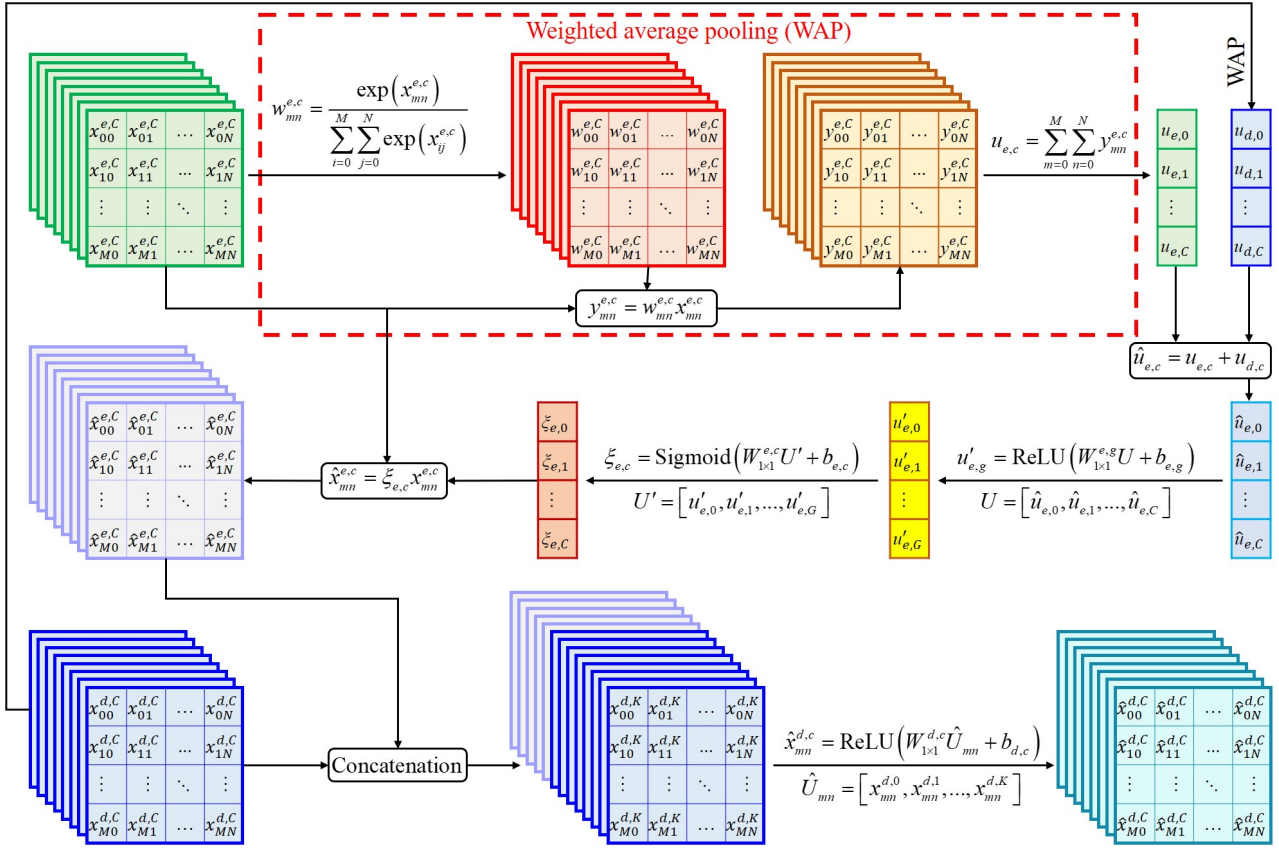


Figure 2: The details of the CAFFM structure.

than the features in the decoder, the multi-channel representations of these shallow features have a high probability of containing irrelevant information, such as background noise. In order to discredit this irrelevant information in the multi-scale features of the feature pyramid before fusing them with the decoder features, we develop a CAFFM (i.e., channel-attention feature fusion module) to guide the skip connection.

The structure of the CAFFM is detailed in Fig. 2. Let  $F_e = \{f_{e,c}\}_{c=0}^C$  be the shallow feature maps from the encoder unit  $e \in \{0, 1, 2, 3\}$ , where each  $f_{e,c} = \left\{ \{x_{mn}^{e,c}\}_{m=0}^M \right\}_{n=0}^N$  represents the feature map in channel  $c \in \{0, \dots, C\}$ .  $x_{mn}^{e,c}$  is the element in  $m$ th ( $m \in \{0, \dots, M\}$ ) row and  $n$ th ( $n \in \{0, \dots, N\}$ ) column of the feature map  $f_{e,c}$ . We first use a weighted average pooling (WAP) method to calculate the channel score  $u_{e,c}$  of each feature map in  $F_e$ , which is formulated as:

$$u_{e,c} = \sum_{m=0}^M \sum_{n=0}^N w_{mn}^{e,c} x_{mn}^{e,c} \quad (1)$$

where  $w_{mn}^{e,c}$  is a self-produced weight coefficient that filters each element  $x_{mn}^{e,c}$  in  $f_{e,c}$ , and it is computed as:

$$w_{mn}^{e,c} = \frac{\exp(x_{mn}^{e,c})}{\sum_{i=0}^M \sum_{j=0}^N \exp(x_{ij}^{e,c})} \quad (2)$$

Similarly, given the up-sampled deep feature maps  $F_d = \{f_{d,c}\}_{c=0}^C$  from the decoder unit  $d \in \{1, 2, 3, 4\}$ , where  $f_{d,c} = \left\{ \left\{ x_{mn}^{d,c} \right\}_{m=0}^M \right\}_{n=0}^N$ , the channel score  $u_{d,c}$  of each feature map in  $F_d$  can also be computed according to Eq.(1) and Eq.(2). Inspired by Attention U-Net [37] where additive attention was used to produce the attention coefficients, the resulting channel scores of shallow and deep feature maps are then summed as:

$$\hat{u}_{e,c} = u_{e,c} + u_{d,c} \quad (3)$$

which leads to a coarse channel-wise attention coefficient vector  $U = [\hat{u}_{e,0}, \hat{u}_{e,1}, \dots, \hat{u}_{e,C}]$ .

To encourage the CAFFM to learn the channel-wise dependencies between multi-channel feature representations, we propose to use a convolutional scaling operation to refine the coarse coefficient vector, which is composed of two steps. The first step is a downscaling operation:

$$u'_{e,g} = \text{ReLU}\left(W_{1 \times 1}^{e,g} U + b_{e,g}\right) \quad (4)$$

where  $u'_{e,g}$  is the  $g$ th element in the downscaled vector  $U' = [u'_{e,0}, u'_{e,1}, \dots, u'_{e,G}]$ ,  $G < C$  ( $G = C/4$  in our work).  $W_{1 \times 1}^{e,g}$  and  $b_{e,g}$  are kernel parameters of a  $1 \times 1$  convolution operation that learns the dependencies between each channel score in  $U$ . The second step is an upscaling operation:

$$\xi_{e,c} = \text{Sigmoid} \left( W_{1 \times 1}^{e,c} U' + b_{e,c} \right) \quad (5)$$

where  $\xi_{e,c}$  is the  $c$ th coefficient in the refined channel-wise attention coefficient vector  $\xi_e = [\xi_{e,0}, \xi_{e,1}, \dots, \xi_{e,C}]$  that filters the multi-channel representations of shallow features and preserves only the channel activation relevant to the target.  $\text{Sigmoid}(\cdot)$  represents the Sigmoid function.  $W_{1 \times 1}^{e,c}$  and  $b_{e,c}$  are kernel parameters of a  $1 \times 1$  convolution operation that expands the vector  $U'$  to match the channel dimension of  $F_e$ , i.e., the feature maps from the  $e$ th encoder unit.

Afterward, the shallow feature maps are multiplied with the refined channel-wise attention coefficient vector, and the resulting filtered feature representations are concatenated with the deep feature maps, followed by a  $1 \times 1$  convolution operation to produce the output of the CAFFM, i.e.,  $F_{\text{CAFFM}}$ . It is formulated as:

$$F_{\text{CAFFM}} = \Phi_{\text{ReLU}} \left( \text{Concat} \left( \xi_e F_e, F_d \right) \right) \quad (6)$$

where  $\Phi_{\text{ReLU}}(\cdot)$  represents the  $1 \times 1$  convolution operation with ReLU activation function that halves the input channels.  $\text{Concat}(\cdot)$  is the concatenation function.

It is worth mentioning that the values of  $e$  are taken opposite to that of  $d$  due to the contracting-to-expanding structure. Theoretically, the global average pooling equally considers all feature pixels and thus is insensitive to the feature category. The max pooling only considers the maximum feature pixel, which causes a loss of relevant information. However, the proposed WAP adaptively considers feature pixels according to the weights of their activation, thus producing channel scores more sensitive to relevant features and, in turn, leads to channel attention more effectively suppressing irrelevant information while keeping all relevant ones in shallow features during the skip connection.

## 4. Experiments and results

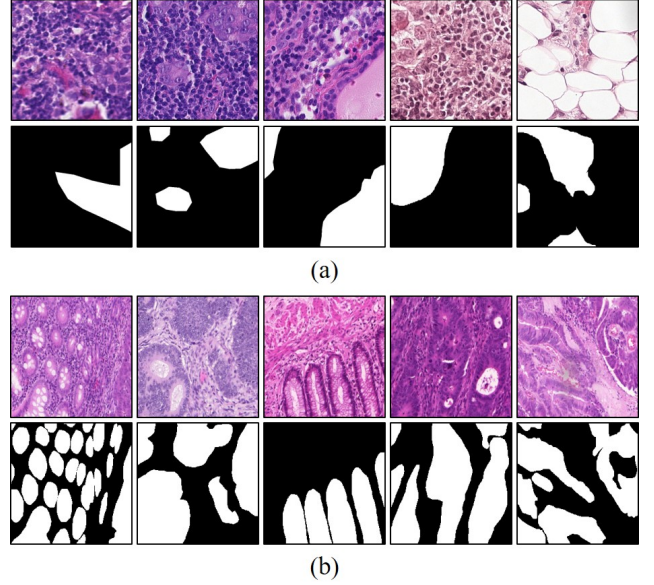
We used two publicly available histopathological image datasets with a large and small scale, respectively, to evaluate the performance of our CA-SegNet. In addition, several relevant state-of-the-art segmentation models were selected as baselines in our comparative experiments, including U-Net [23], SegNet [22], UNet++ [31], R2U-Net [32], Attention U-Net [37], MultiResUNet [33], DoubleU-Net [34], R2AU-Net [39], ComBiNet (containing ComBiNetS, ComBiNetM, and ComBiNetL) [42], FANet (containing FANet18 and FANet34) [24], and UCTransNet [43].

### 4.1. Datasets

#### 4.1.1. Camelyon16 patch-based dataset

This dataset (large-scale) is produced from the Camelyon16 dataset [44] based on the protocol proposed in [45].

It is a patch-based dataset cropped from 159 whole-slide metastatic breast cancer images with a size of  $65000 \times 45000$  pixels. The resulting dataset has about 12532 patches for training, 4429 patches for validation, and 7639 patches for testing. All the patches with a size of  $512 \times 512$  pixels have pixel-level annotations (i.e., binary masks) produced by experts. Fig. 3a provides some examples of the images and their corresponding pixel-level annotations in the Camelyon16 patch-based dataset.



**Figure 3:** Examples of images in two datasets. (a) The Camelyon16 patch-based dataset. (b) The GlaS dataset. The upper row is original samples, and the bottom row is pixel-level annotations.

#### 4.1.2. GlaS dataset

It is a dataset (small-scale) of colon cancer images (with large variations in gland shape and size) with a size of  $775 \times 522$  pixels obtained from hematoxylin and eosin (H&E)-stained histology sections [46]. This dataset contains 85 images (20% of which were assigned to the validation set for 5-fold cross-validation) for training and 80 images for testing. All the images have both image-level (i.e., benign or malignant) and pixel-level annotations. Some examples of the GlaS dataset are shown in Fig. 3b.

## 4.2. Experimental settings

### 4.2.1. Evaluation metrics

To thoroughly measure the similarity between the predicted mask and the ground truth for our CA-SegNet and existing state-of-the-art methods, three commonly used evaluation metrics were calculated in the comparative experiments, including the mean intersection over union (mIoU), mean dice coefficient (mDice), and mean precision (mPrecision). Furthermore, the pixel-level precision-recall (P-R) curve was computed.



#### 4.2.2. Implementation details

To demonstrate the effectiveness of the proposed CA-SegNet architecture and make fair comparisons with existing state-of-the-art methods, the commonly used binary cross-entropy loss function was selected for all networks. The Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and a weight decay of  $1 \times 10^{-4}$  was used to train networks. Inspired by the work in [47], we froze the first convolutional layer and fine-tuned all the pre-trained layers with a learning rate initialized at  $1 \times 10^{-5}$  for the transfer learning-based methods, i.e., the proposed CA-SegNet, SegNet, DoubleU-Net, and FANet. For network layers without pre-training and other networks, the learning rate was initialized at  $1 \times 10^{-3}$ . We set a batch size of 32 and a training epoch of 60 for the Camelyon16 patch-based dataset while a batch size of 8 and a training epoch of 500 for the GlaS dataset. All the images were resized to  $256 \times 256$  pixels before input to networks. In addition, random rotation with an angle in  $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ , random vertical and horizontal flips with a probability of 0.5, and random color jittering (brightness = 0.5, contrast = 0.5, saturation = 0.5, and hue = 0.05) were performed to each image for data augmentation. Both datasets were trained with 5-fold cross-validation, and the final results were obtained through the average of the best results from these five groups of experiments.

#### 4.3. Results on the Camelyon16 patch-based dataset

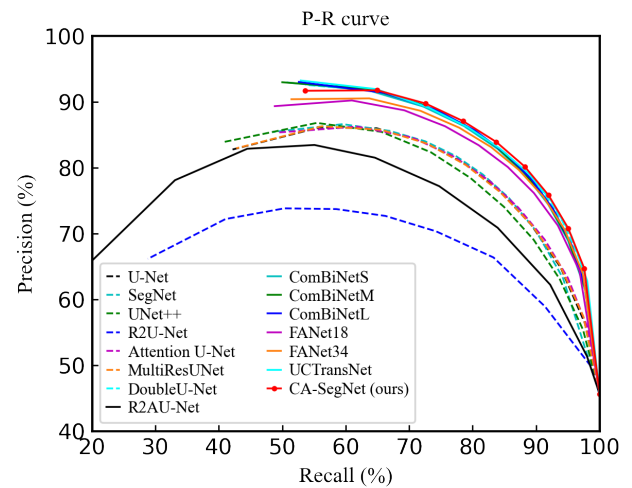
The quantitative comparison results of the proposed CA-SegNet and existing state-of-the-art methods on the Camelyon16 patch-based dataset are presented in Table 3. It shows that our CA-SegNet has achieved the best performance with approximately 72.14% in mIoU, 81.49% in mDice, and 83.95% in mPrecision among all the segmentation models. Compared to the U-Net series, i.e., U-Net, UNet++, R2U-Net, Attention U-Net, MultiResUNet, DoubleU-Net, and R2AU-Net, our CA-SegNet achieves an improvement of 0.64% to 10.82% in mIoU, 0.42% to 9.36% in mDice, and 0.48% to 5.39% in mPrecision. Compared to attention-based CNNs, i.e., Attention U-Net, R2AU-Net, FANet18 and FANet34, the CA-SegNet achieves an improvement of at least 1.15% in mIoU, 1.01% in mDice, and 0.88% in mPrecision. Finally, compared to the transformer-based UCTransNet, our CA-SegNet leads to an improvement of 0.30% in mIoU, 0.11% in mDice, and 0.56% in mPrecision, suggesting the competitive performance of our CA-SegNet against the transformer-based method in terms of histopathological image segmentation. The same results can be observed in Fig. 4, where the P-R curve of our CA-SegNet is more toward the top right corner, indicating better performance.

The visual comparison results of different methods on the Camelyon16 patch-based dataset are shown in Fig. 5, where we can observe that our CA-SegNet has a more powerful ability to identify metastatic tissues and produces the segmentation prediction closer to the ground truth. Note

**Table 3**

Quantitative comparisons to existing state-of-the-art networks on the Camelyon16 patch-based dataset. The best and second-best results are marked in red and blue, respectively.

Methods	mIoU	mDice	mPrecision
U-Net [23]	67.09±0.51	76.91±0.55	81.64±0.33
SegNet [22]	67.63±0.29	77.25±0.32	81.62±0.28
UNet++ [31]	66.03±0.47	76.11±0.46	81.14±0.27
R2U-Net [32]	62.79±1.16	73.85±1.03	78.56±0.61
Attention U-Net [37]	67.09±0.37	76.99±0.39	81.46±0.20
MultiResUNet [33]	67.38±0.43	77.31±0.47	81.24±0.19
DoubleU-Net [34]	71.50±0.36	81.07±0.25	<b>83.47±0.24</b>
R2AU-Net [39]	61.32±0.55	72.13±0.49	78.75±0.61
ComBiNetS [42]	71.42±0.29	81.14±0.25	82.79±0.16
ComBiNetM [42]	71.61±0.23	81.30±0.17	82.90±0.30
ComBiNetL [42]	71.59±0.22	81.23±0.15	83.10±0.29
FANet18 [24]	70.14±0.39	79.79±0.35	82.60±0.21
FANet34 [24]	70.99±0.41	80.48±0.42	83.07±0.13
UCTransNet [43]	<b>71.84±0.05</b>	<b>81.38±0.06</b>	83.39±0.21
CA-SegNet (ours)	<b>72.14±0.18</b>	<b>81.49±0.19</b>	<b>83.95±0.34</b>



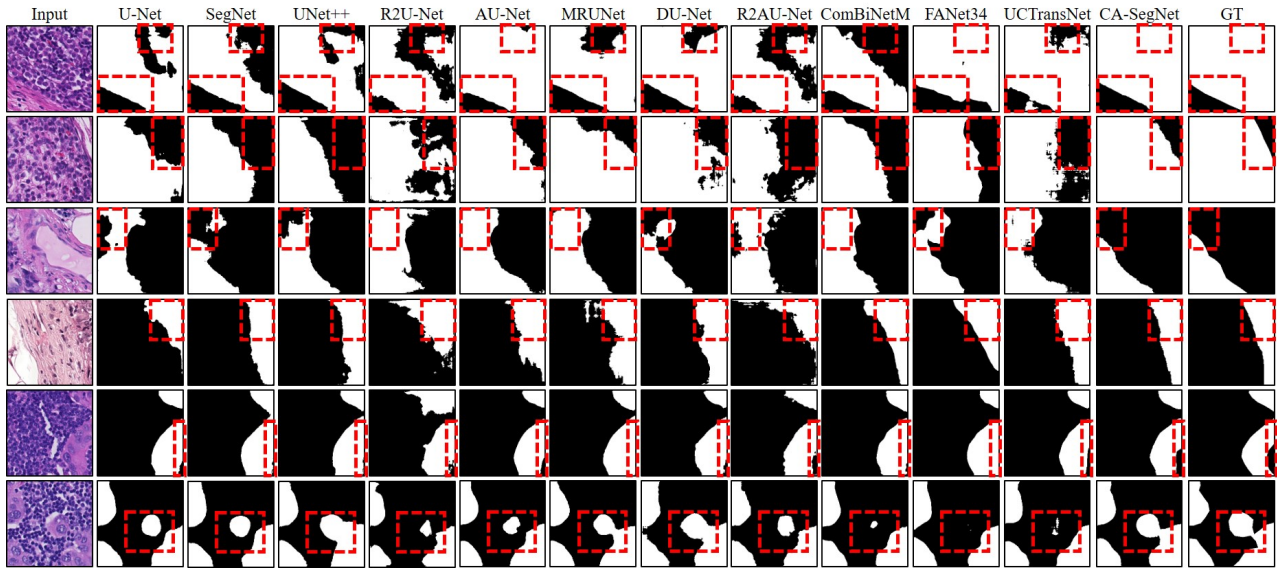
**Figure 4:** P-R curve of different methods on the Camelyon16 patch-based dataset.

that only the best versions of ComBiNet (i.e., ComBiNetM) and FANet (i.e., FANet34) are presented. In addition, it can be seen from rows 2, 3, and 5 that our CA-SegNet is more precise in dealing with target boundaries. The CA-SegNet also keeps better integrity of the target tissues, as seen from rows 1 and 6.

#### 4.4. Results on the GlaS dataset

To evaluate the effectiveness and demonstrate the superior performance of our CA-SegNet on small-scale histopathological image datasets, all the comparative networks, including the CA-SegNet, were performed on the GlaS dataset.

The quantitative comparison results are given in Table 4. We can observe that our CA-SegNet outperforms all the state-of-the-art methods with approximately 85.06% in mIoU, 91.43% in mDice, and 91.78% in mPrecision, which



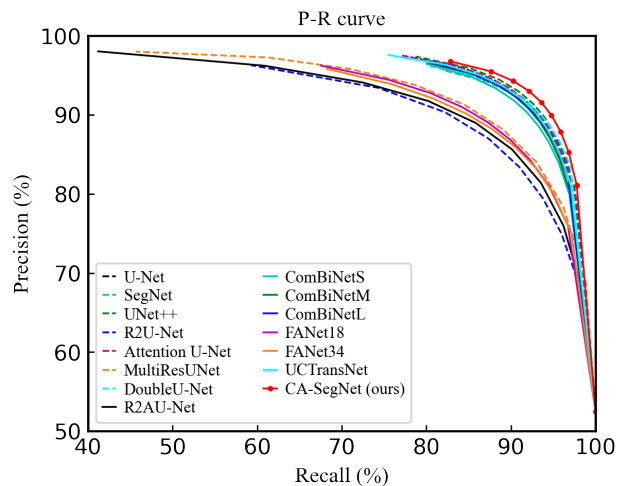
**Figure 5:** Segmentation results of images in the Camelyon16 patch-based dataset. Regions highlighted by the dashed boxes are for better visualization of the prediction differences. AU-Net, MRUNet, DU-Net and GT represent Attention U-Net, MultiResUNet, DoubleU-Net and the ground truth, respectively.

**Table 4**

Quantitative comparisons to existing state-of-the-art networks on the GlaS dataset. The best and second-best results are marked in red and blue, respectively.

Methods	mIoU	mDice	mPrecision
U-Net [23]	83.73±0.40	90.51±0.30	90.84±0.20
SegNet [22]	83.32±0.45	90.32±0.34	90.72±0.30
UNet++ [31]	<b>84.08±0.21</b>	<b>90.77±0.10</b>	<b>90.96±0.26</b>
R2U-Net [32]	75.99±0.54	85.51±0.34	86.32±0.45
Attention U-Net [37]	83.58±0.32	90.38±0.26	90.80±0.17
MultiResUNet [33]	77.98±0.78	86.71±0.47	87.24±0.38
DoubleU-Net [34]	83.58±0.28	90.58±0.18	90.70±0.28
R2AU-Net [39]	76.52±1.37	85.79±1.02	86.79±0.59
ComBiNetS [42]	82.51±0.53	89.71±0.40	89.75±0.31
ComBiNetM [42]	83.19±0.25	90.18±0.19	90.27±0.24
ComBiNetL [42]	83.19±0.23	90.22±0.14	90.24±0.21
FANet18 [24]	77.60±0.28	86.68±0.16	86.95±0.24
FANet34 [24]	77.19±0.37	86.35±0.24	86.56±0.33
UCTransNet [43]	83.40±0.32	90.24±0.20	90.38±0.27
CA-SegNet (ours)	<b>85.06±0.22</b>	<b>91.43±0.17</b>	<b>91.78±0.19</b>

indicates the better robustness of the CA-SegNet to small-scale datasets. Some subsequent variants of the general U-Net, i.e., MultiResUNet, DoubleU-Net, ComBiNet, FANet, and UCTransNet, perform worse than the U-Net, which is opposite to that on the Camelyon16 patch-based dataset, suggesting their strong dependencies on the dataset scales. Furthermore, our CA-SegNet leads to an improvement of approximately 0.98% in mIoU, 0.66% in mDice, and 0.82% in mPrecision compared to the second-best network UNet++. The P-R curves of our CA-SegNet and other networks on the GlaS dataset are shown in Fig. 6. Similarly, CA-SegNet is always toward the top right corner. Fig. 7 provides the visual comparison of different networks on the GlaS dataset. It is



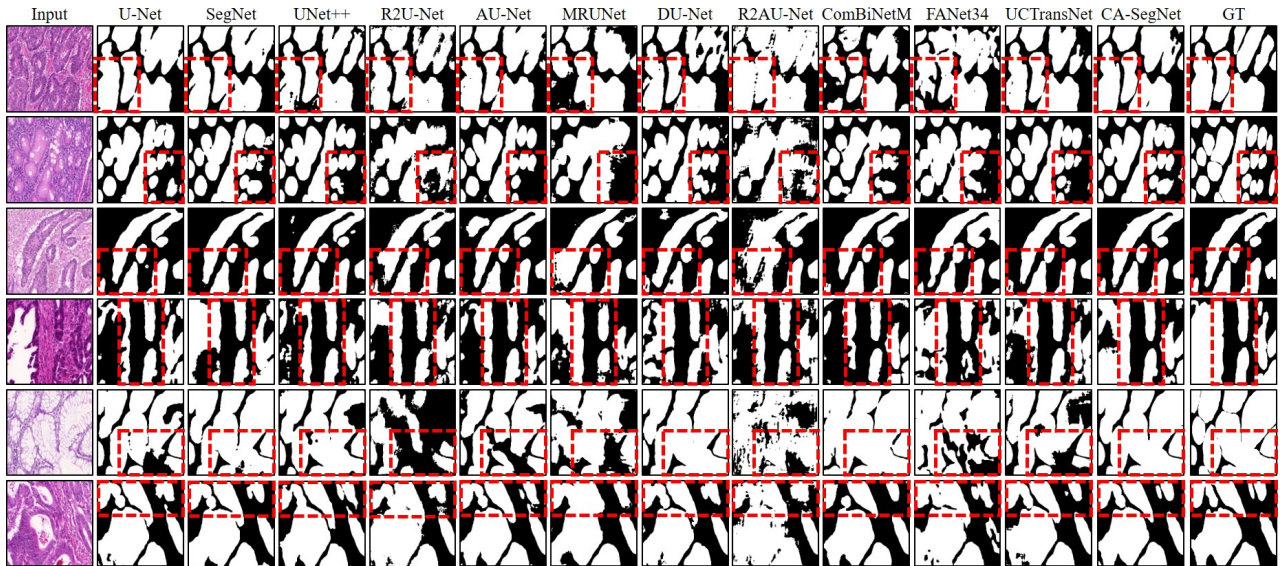
**Figure 6:** P-R curve of different methods on the GlaS dataset.

observed that our CA-SegNet still achieves the segmentation results closest to the ground truth.

## 5. Discussion

### 5.1. Ablation study

To analyze the individual contribution of the key components (i.e., bottleneck encoder and CAFFM) to our CA-SegNet in histopathological image segmentation. We conducted the ablation study on the two datasets. We started with the standard encoder-decoder segmentation structure (E-SD) without skip connection and with the encoder and decoder symmetrical to each other. Afterward, the standard decoder of the E-SD was modified to the bottleneck decoder, and this structure is named E-BD. The basic FFM that



**Figure 7:** Segmentation examples of images in the GlaS dataset. Regions highlighted by the dashed boxes are for better visualization of the prediction differences. AU-Net, MRUNet, DU-Net and GT represent Attention U-Net, MultiResUNet, DoubleU-Net and the ground truth, respectively.

achieves the skip connection in U-Net was then added to the E-BD structure (E-BD+FFM). Finally, the FFM was replaced with the CAFFM, which constructs the proposed CA-SegNet, i.e., E-BD+CAFFM.

**Table 5**

Quantitative segmentation results on two datasets, i.e., the Camelyon16 patch-based (CPB) dataset and the GlaS dataset. The best results are highlighted in bold.

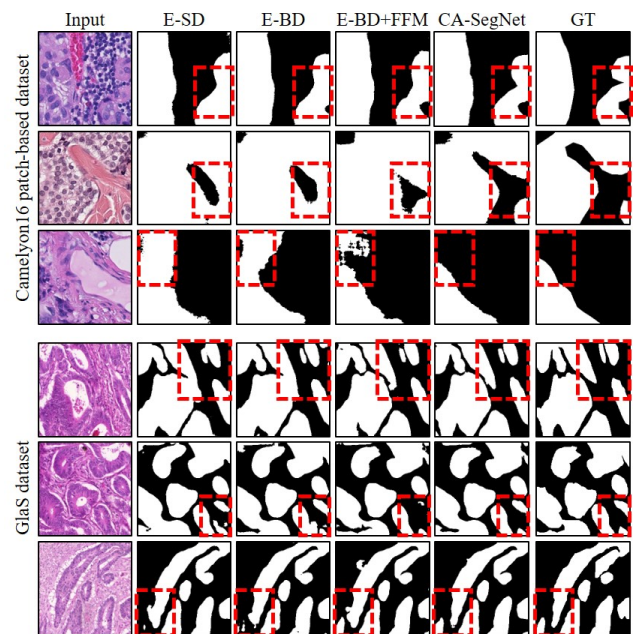
Dataset	Methods	mIoU	mDice	mPrecision
CPB	E-SD	68.04±0.42	77.57±0.45	82.01±0.29
	E-BD	68.45±0.50	77.95±0.55	81.99±0.29
	E-BD+FFM	68.45±0.44	77.96±0.41	82.14±0.32
	CA-SegNet	<b>72.14±0.18</b>	<b>81.49±0.19</b>	<b>83.95±0.34</b>
GlaS	E-SD	83.44±0.53	90.43±0.35	90.82±0.38
	E-BD	83.71±0.40	90.53±0.32	91.00±0.17
	E-BD+FFM	84.81±0.31	91.24±0.23	91.68±0.20
	CA-SegNet	<b>85.06±0.22</b>	<b>91.43±0.17</b>	<b>91.78±0.19</b>

**Table 6**

Computational complexity of structures with different components in ablation study

Methods	Parameters (M)	MACs (G)
E-SD	34.94	50.40
E-BD	20.74	32.97
E-BD+FFM	21.44	35.15
CA-SegNet	21.62	35.15

The quantitative segmentation results and computational complexity of structures with different components are provided in Table 5 and Table 6, respectively. The segmentation performance improves with the addition (or modification)



**Figure 8:** Segmentation examples of images in the Camelyon16 patch-based dataset (upper) and the GlaS dataset (bottom) in the ablation study. GT represents the ground truth.

of each component. More specifically, the E-BD structure exhibits an improvement of approximately 0.41% in mIoU on the Camelyon16 patch-based dataset and 0.27% in mIoU on the GlaS dataset compared to E-SD, as well as distinctive decreases in parameters and MACs, suggesting the better feature integration ability and lower computational complexity of our bottleneck decoder. Compared with the E-BD structure, the additional CAFFM in our CA-SegNet achieves an improvement of 1.35% to 3.69% in mIoU, 0.90%

to 3.54% in mDice, and 0.78% to 1.96% in mPrecision on the two datasets while having a slight increase in computational complexity. In addition, compared with the basic FFM of the E-BD+FFM structure, the CAFFM reaches an improvement of up to 3.69% in mIoU, 3.53% in mDice, and 1.81% in mPrecision with the same MACs. We believe these improvements are gained owing to the fact that our channel-wise attention coefficients filter undesired information in multi-channel representations of the shallow features. The visual results of the ablation study on two datasets are shown in Fig. 8, which further confirms the effectiveness of each proposed component. It is observed that the segmentation results approach the ground truth with the addition of our bottleneck decoder and CAFFM, such as the first and third rows.

### 5.2. Interpolation vs. max-unpooling

Apart from the ablation study in the above section studying the proposed key components, we also conducted a comparative experiment between the commonly used interpolation (i.e., bilinear) up-sampling operation and the max-unpooling operation to investigate their effects on the segmentation performance. The evaluation results are given in Table 7. We can observe that the max-unpooling operation yields better results in terms of all three evaluation metrics on both datasets. We believe this superior performance benefits from the saved locations of the maximum feature values that define targets.

**Table 7**

Performance of CA-SegNet based on different up-sampling methods on two datasets, i.e., the Camelyon16 patch-based (CPB) dataset and the GlaS dataset. The best results are highlighted in bold.

Dataset	Methods	mIoU	mDice	mPrecision
CPB	Interpolation	71.74±0.29	81.17±0.27	83.81±0.30
	Max-unpooling	<b>72.14±0.18</b>	<b>81.49±0.19</b>	<b>83.95±0.34</b>
GlaS	Interpolation	84.37±0.27	91.01±0.14	91.39±0.10
	Max-unpooling	<b>85.06±0.22</b>	<b>91.43±0.17</b>	<b>91.78±0.19</b>

### 5.3. Effect of image resolution

We conducted a comparative experiment involving three image resolutions, i.e., 128×128 pixels, 256×256 pixels, and 512×512 pixels, to study the effect of the input resolution on the performance of our CA-SegNet. Table 8 provides the evaluation results, where we can observe that the segmentation performance changes according to input resolutions. Interestingly, the 256×256 case achieves the best results compared to the other two resolution cases, suggesting that higher input resolutions do not necessarily produce better performance for the proposed CA-SegNet. We believe this is because a fixed network structure with a constant receptive field determines the limit of the ability to capture contextual semantic information of input images.

**Table 8**

Performance of CA-SegNet with input images of different resolutions on two datasets, i.e., the Camelyon16 patch-based (CPB) dataset and the GlaS dataset. The best results are highlighted in bold.

Dataset	Input Resolution	mIoU	mDice	mPrecision
CPB	128×128	72.14±0.19	81.40±0.15	83.79±0.19
	256×256	<b>72.14±0.18</b>	<b>81.49±0.19</b>	<b>83.95±0.34</b>
	512×512	71.37±0.19	81.00±0.20	83.50±0.14
GlaS	128×128	81.31±0.17	88.93±0.13	89.37±0.26
	256×256	<b>85.06±0.22</b>	<b>91.43±0.17</b>	<b>91.78±0.19</b>
	512×512	85.04±0.19	91.38±0.11	91.75±0.10

### 5.4. Effect of network depth

In this section, we conducted a comparative experiment to study the effect of network depth on segmentation performance. In addition to VGG16, the other networks based on VGG11 and VGG19 were constructed as the shallower and deeper versions of our CA-SegNet. Note that the decoder layers of these two versions were changed accordingly. The comparison results are provided in Table 9. It is shown that the performance improves with the addition of the network layers on the Camelyon16 patch-based dataset, while the VGG16-based CA-SegNet achieves the best results (except for the mPrecision) on the GlaS dataset. We believe this is because the small-scale dataset GlaS leads to the overfitting of the deeper network.

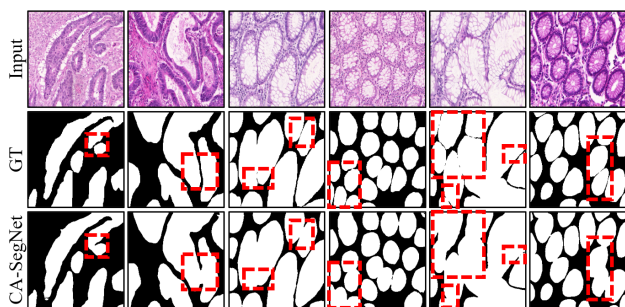
**Table 9**

Performance of CA-SegNet with different network depths on two datasets, i.e., the Camelyon16 patch-based (CPB) dataset and the GlaS dataset. The best results are highlighted in bold.

Dataset	Methods	mIoU	mDice	mPrecision
CPB	VGG11-based	71.81±0.35	81.26±0.34	83.62±0.32
	VGG16-based	72.14±0.18	81.49±0.19	83.95±0.34
	VGG19-based	<b>72.24±0.15</b>	<b>81.53±0.09</b>	<b>84.05±0.43</b>
GlaS	VGG11-based	84.73±0.36	91.23±0.27	91.47±0.27
	VGG16-based	<b>85.06±0.22</b>	<b>91.43±0.17</b>	91.78±0.19
	VGG19-based	84.92±0.25	91.29±0.18	<b>91.82±0.19</b>

### 5.5. Limitation analysis

Despite the outstanding performance of our CA-SegNet compared to existing segmentation models, there is a limitation in segmenting multiple objects with tiny gaps. As seen in Fig. 9, our CA-SegNet fails to identify boundaries of some tightly distributed glands, which may be caused by the lack of data volume and diversity in the GlaS dataset. The failure to retain relevant information from shallow layers during skip connection may be another reason.



**Figure 9:** Segmentation examples with limitation in the GlaS dataset.

## 6. Conclusion

We proposed a channel-attention encoder-decoder architecture named CA-SegNet to segment histopathological images. Our CA-SegNet was built on top of the final convolutional layer of the ImageNet-trained VGG16 model to avoid network overfitting and reduce convergence duration, where the pre-trained layers form the encoder. We reconsidered the standard decoder structure and developed a bottleneck-structured decoder to integrate relevant contextual information in multi-level features from the encoder more effectively. In addition, we designed a sequence of CAFFMs in skip connections between the encoder and decoder to eliminate irrelevant information (e.g., background noise) within multi-channel feature representations from shallow layers, where a WAP method was proposed to compute the channel scores to produce attention coefficients and a convolutional scaling operation was adopted to learn the channel-wise dependencies. Extensive experiments on two public histopathological image datasets demonstrated the effectiveness of the bottleneck decoder and CAFFM and the superior performance of our CA-SegNet compared to existing state-of-the-art segmentation methods. For future work, it would be interesting to incorporate more attention mechanisms into our network to avoid limitations and consider the network for more medical image modalities.

## CRedit authorship contribution statement

**Feng He:** Conceptualization, Data curation, Methodology, Software, Validation, Writing - original draft. **Weibo Wang:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing. **Lijuan Ren:** Conceptualization, Software, Visualization, Writing - review & editing. **Yixuan Zhao:** Conceptualization, Data curation, Validation, Writing - review & editing. **Zhengjun Liu:** Conceptualization, Project administration, Writing - review & editing. **Yuemin Zhu:** Conceptualization, Methodology, Project administration, Supervision, Validation, Writing - review & editing.

## 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 in part by the National Natural Science Foundation of China (52275527, 51975161 and 52275526), the Key Research and Development Program of Heilongjiang (Grant No. 2022ZX01A27), CGN-HIT Advanced Nuclear and New Energy Research Institute (CGN-HIT202201), in part by the International Research Project METISLAB, and in part by the China Scholarship Council.

## References

- [1] B. Xu, J. Liu, X. Hou, B. Liu, J. Garibaldi, I. O. Ellis, A. Green, L. Shen, G. Qiu, Attention by selection: A deep selective attention approach to breast cancer classification, *IEEE transactions on medical imaging* 39 (2019) 1930–1941.
- [2] R. Yan, F. Ren, Z. Wang, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, F. Zhang, Breast cancer histopathological image classification using a hybrid deep neural network, *Methods* 173 (2020) 52–60.
- [3] K. Ding, M. Zhou, H. Wang, O. Gevaert, D. Metaxas, S. Zhang, A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer, *Scientific Data* 10 (2023) 231.
- [4] A. B. Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, G. Forestier, C. Wemmert, Deep learning for colon cancer histopathological images analysis, *Computers in Biology and Medicine* 136 (2021) 104730.
- [5] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, *IEEE Transactions on Instrumentation and Measurement* 71 (2022) 1–15.
- [6] X. Li, S. Pang, R. Zhang, J. Zhu, X. Fu, Y. Tian, J. Gao, Attransunet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation, *Computers in Biology and Medicine* 152 (2023) 106365.
- [7] Z. Li, J. Zhang, T. Tan, X. Teng, X. Sun, H. Zhao, L. Liu, Y. Xiao, B. Lee, Y. Li, et al., Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the acdc@ lunghp challenge 2019, *IEEE Journal of Biomedical and Health Informatics* 25 (2020) 429–440.
- [8] W. Shao, T. Wang, Z. Huang, Z. Han, J. Zhang, K. Huang, Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images, *IEEE Transactions on Medical Imaging* 40 (2021) 3739–3747.
- [9] H.-Y. Chiu, H.-S. Chao, Y.-M. Chen, Application of artificial intelligence in lung cancer, *Cancers* 14 (2022) 1370.
- [10] H. Tokunaga, Y. Teramoto, A. Yoshizawa, R. Bise, Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12597–12606.
- [11] R. Schmitz, F. Madesta, M. Nielsen, J. Krause, S. Steurer, R. Werner, T. Rösch, Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture, *Medical image analysis* 70 (2021) 101996.
- [12] C. Nguyen, Z. Asad, R. Deng, Y. Huo, Evaluating transformer-based semantic segmentation networks for pathological image segmentation, in: *Medical Imaging 2022: Image Processing*, volume 12032, SPIE, 2022, pp. 942–947.
- [13] P. Sharma, A. Gautam, P. Maji, R. B. Pachori, B. K. Balabantaray, Lisegpnnet: Encoder-decoder mode lightweight segmentation network for colorectal polyps analysis, *IEEE Transactions on Biomedical Engineering* 70 (2022) 1330–1339.

- [14] T. Rasal, T. Veerakumar, B. N. Subudhi, S. Esakirajan, Segmentation of gastric cancer from microscopic biopsy images using deep learning approach, *Biomedical Signal Processing and Control* 86 (2023) 105250.
- [15] S. Khare, A. Nishad, A. Upadhyay, V. Bajaj, Classification of emotions from eeg signals using time-order representation based on the s-transform and convolutional neural network, *Electronics Letters* 56 (2020) 1359–1361.
- [16] M. Byra, Breast mass classification with transfer learning based on scaling of deep representations, *Biomedical Signal Processing and Control* 69 (2021) 102828.
- [17] Z. Gao, Z. Lu, J. Wang, S. Ying, J. Shi, A convolutional neural network and graph convolutional network based framework for classification of breast histopathological images, *IEEE Journal of Biomedical and Health Informatics* 26 (2022) 3163–3173.
- [18] N. Phukan, M. S. Manikandan, R. B. Pachori, Afibri-net: A lightweight convolution neural network based atrial fibrillation detector, *IEEE Transactions on Circuits and Systems I: Regular Papers* (2023).
- [19] D. Das, D. R. Nayak, R. B. Pachori, Ca-net: A novel cascaded attention-based network for multi-stage glaucoma classification using fundus images, *IEEE Transactions on Instrumentation and Measurement* (2023).
- [20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
- [21] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [22] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE transactions on pattern analysis and machine intelligence* 39 (2017) 2481–2495.
- [23] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18, Springer, 2015, pp. 234–241.
- [24] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, S. Sclaroff, Real-time semantic segmentation with fast attention, *IEEE Robotics and Automation Letters* 6 (2020) 263–270.
- [25] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, H. Lu, Resgnet: Residual group attention network for medical image classification and segmentation, *Medical Image Analysis* 76 (2022) 102313.
- [26] Y. Fang, H. Huang, W. Yang, X. Xu, W. Jiang, X. Lai, Nonlocal convolutional block attention module vnet for gliomas automatic segmentation, *International Journal of Imaging Systems and Technology* 32 (2022) 528–543.
- [27] A. Chakravarty, J. Sivaswamy, Race-net: a recurrent neural network for biomedical image segmentation, *IEEE journal of biomedical and health informatics* 23 (2018) 1151–1162.
- [28] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, D. Shen, High-resolution encoder-decoder networks for low-contrast medical image segmentation, *IEEE Transactions on Image Processing* 29 (2019) 461–475.
- [29] J. M. J. Valanarasu, V. M. Patel, Unext: Mlp-based rapid medical image segmentation network, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 23–33.
- [30] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.
- [31] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 3–11.
- [32] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, V. K. Asari, Recurrent residual u-net for medical image segmentation, *Journal of Medical Imaging* 6 (2019) 014006–014006.
- [33] N. Ibtihaz, M. S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, *Neural networks* 121 (2020) 74–87.
- [34] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, H. D. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, IEEE, 2020, pp. 558–564.
- [35] Z. Ning, S. Zhong, Q. Feng, W. Chen, Y. Zhang, Smu-net: Saliency-guided morphology-aware u-net for breast lesion segmentation in ultrasound image, *IEEE transactions on medical imaging* 41 (2021) 476–490.
- [36] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, V. M. Patel, Kiunet: Overcomplete convolutional architectures for biomedical image and volumetric segmentation, *IEEE Transactions on Medical Imaging* 41 (2021) 965–976.
- [37] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, *Medical image analysis* 53 (2019) 197–207.
- [38] E. Thomas, S. Pawan, S. Kumar, A. Horo, S. Niyas, S. Vinayagamani, C. Kesavadas, J. Rajan, Multi-res-attention unet: a cnn model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images, *IEEE Journal of Biomedical and Health Informatics* 25 (2020) 1724–1734.
- [39] Q. Zuo, S. Chen, Z. Wang, R2au-net: attention recurrent residual convolutional neural network for multimodal medical image segmentation, *Security and Communication Networks* 2021 (2021) 1–10.
- [40] Y. Yuan, L. Zhang, L. Wang, H. Huang, Multi-level attention network for retinal vessel segmentation, *IEEE Journal of Biomedical and Health Informatics* 26 (2021) 312–323.
- [41] H. Yin, Y. Shao, Cfu-net: A coarse-fine u-net with multi-level attention for medical image segmentation, *IEEE Transactions on Instrumentation and Measurement* (2023).
- [42] M. Ferianc, D. Manocha, H. Fan, M. Rodrigues, Combinet: Compact convolutional bayesian neural network for image segmentation, in: *Artificial Neural Networks and Machine Learning—ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part III* 30, Springer, 2021, pp. 483–494.
- [43] H. Wang, P. Cao, J. Wang, O. R. Zaiane, Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2022, pp. 2441–2449.
- [44] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, *Jama* 318 (2017) 2199–2210.
- [45] J. Rony, S. Belharbi, J. Dolz, I. B. Ayed, L. McCaffrey, E. Granger, Deep weakly-supervised learning methods for classification and localization in histology images: a survey, *arXiv preprint arXiv:1909.03354* (2019).
- [46] K. Sirinukunwattana, D. R. Snead, N. M. Rajpoot, A stochastic polygons model for glandular structures in colon histology images, *IEEE transactions on medical imaging* 34 (2015) 2366–2378.
- [47] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, K. H. Cha, Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets, *IEEE transactions on medical imaging* 38 (2018) 686–696.