IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DEEP LEARNING FOR COMPUTER-AIDED MEDICAL DIAGNOSIS

Received April 27, 2019, accepted May 13, 2019, date of publication May 24, 2019, date of current version June 21, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2918800*

# Accurate Gastric Cancer Segmentation in Digital Pathology Images Using Deformable Convolution and Multi-Scale Embedding Networks

**MUYI SUN[1,2], GUANHONG ZHANG[1,2], HAO DANG[1,2], XINGQUN QI[1,2], XIAOGUANG ZHOU[1,2], AND QING CHANG[3,4]**

[1]School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Engineering Research Center of Information Network, Ministry of Education, Beijing 100876, China
[3]Shanghai General Practice Medical Education and Research Center, Shanghai 201800, China
[4]Jiading District Central Hospital Affiliated Shanghai University of Medicine & Health Sciences, Shanghai 201800, China

Corresponding authors: Xiaoguang Zhou (zxg@bupt.edu.cn) and Qing Chang (robie0510@hotmail.com)

This work was supported in part by the Open Foundation of State key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, under Grant SKLNST-2018-1-18.

**ABSTRACT** Automatic gastric cancer segmentation is a challenging problem in digital pathology image analysis. Accurate segmentation of gastric cancer regions can efficiently facilitate clinical diagnosis and pathological research. Technically, this problem suffers from various sizes, vague boundaries, and the non-rigid characters of cancerous regions. For addressing these challenges, we use a deep learning based method and integrate several customized modules. Structurally, we replace the basic form of convolution with deformable and Atrous convolutions in specific layers, for adapting to the non-rigid characters and larger receptive field. We take advantage of the Atrous Spatial Pyramid Pooling module and encoder-decoder based semantic-level embedding networks for multi-scale segmentation. In addition, we propose a lightweight decoder to fuse the contexture information, and utilize the dense upsampling convolution for boundary refinement at the end of the decoder. Experimentally, sufficient comparative experiments are enforced on our own gastric cancer segmentation dataset, which is delicately annotated to pixel-level by medical specialists. The quantitative comparisons against several prior methods demonstrate the superiority of our approach. We achieve 91.60% for pixel-level accuracy and 82.65% for mean Intersection over Union.

**INDEX TERMS** Digital pathology image analysis, deformable convolution, dense upsampling convolution, gastric cancer segmentation, multi-scale embedding.

## I. INTRODUCTION

Gastric cancer is one of the most lethal cancers in the world. In 2018, gastric cancer is responsible for over 1,000,000 new cases and an estimated 783,000 deaths. This quantitative analysis makes it the fifth most frequently diagnosed cancer and the third leading cause of cancer death [1]. Clinical pathological analysis makes sense for diagnosis of this cancer and brings the most significant information on the focus of infection. Diagnosing gastric cancer is hard and time-consuming due to image high resolution [1]. In addition, on account of the different diagnostic criteria of different clinicians, an accurate and timely diagnosis becomes luxurious for patients. With repeated diagnoses, the golden time for the treatment might be missed. Therefore, an automatically accurate segmentation method for gastric cancer is indispensable.

With the rapid development of deep learning in digital pathology image analysis, research about pathological segmentation methods has attracted much attention in recent years. Through the creation of datasets [2] and the implementation of algorithms. Years of efforts have launched research into application fields ranging from cell-level to tissue-level [3], [4]. Aïcha and Ghassan [4] have proposed topology aware fully convolutional networks for histology gland segmentation, in which they have considered high-level shape priors and designed a topology-aware loss for training. Bi et al. [5] have provided stacked fully convolutional networks for cell segmentation of rectal cancer. Zhang et al. [6]

The associate editor coordinating the review of this manuscript and approving it for publication was Shuihua Wang.

have designed a framework of adversarial networks for tissue segmentation utilizing unannotated images. In order to optimally use the limited data, Yang et al. [7] have used a deep active learning framework for biomedical image segmentation, including pathological tissues. In relevant research, image-patch based deep learning methods [8]–[10] also have been applied for segmentation or identification on pathological images. In [8], Wang and Khosla have used overlapping image patches for detection of cancer metastasis. In [9], Coudray and Moraria have applied image classification networks for classification and mutation prediction in Non-small cell lung cancer histopathology images. Li et al. [10] have provided a similar approach in gastric cancer identification. However, there are still many challenges in pathological segmentation task, such as the rarity of professional pathological image datasets and the complexity of feature representation in various images. Most previous research has focused on applying deep learning methods to solve pathological problems, ignoring more detailed aspects on the images, such as various sizes, vague boundaries and the non-rigid characters, which are illustrated in Fig. 1. Other studies [11], [12] have also been proposed to address these problems. In [11], Ronneberger and Fischer have proposed an encoder-decoder architecture with skip-connections to embed feature maps from different semantic levels. In [12], a similar method has been used on patch based histological segmentation. Though these methods have shown some advantages for multi-scale perception and got better boundaries, their architectures might not be specifically designed to these problems and the performance might not be satisfactory in our task as shown in Figure. 7(c).

Motivated by the above observations, we have proposed a holistic solution. Firstly, we created a gastric cancer segmentation dataset. Then we have gained the annotations by medical specialists two times per image for annotation and refinement. For accurate segmentation, we have paid more attention to adjusting the receptive field of networks. Receptive field plays a decisive role in feature perception in the field of semantic segmentation. There are two ways commonly employed on changing the receptive field, enlarging the rectangle receptive field [13] and deforming the kernels of operations [14]–[16]. For enlarging, applying cascaded convolutional layers and pooling layers [13] is a widely accepted thought. Dilated convolution [17] is an alternative choice which has been mostly used in recent years [18]–[20]. We have integrated the deep neural networks with dilated convolution and deformable convolution as our basic skeleton frame. We have utilized the ResNet [21], [22] based encoder-decoder network, in which both encoder and decoder have been finely designed. In order to learn the features of different scales, we have proposed a multi-scale embedding strategy. In this strategy, Atrous Spatial Pyramid Pooling module [23] and encoder-decoder based semantic-level embedding networks [11], [24]–[27] have been integrated.

During the clinical diagnosis, a relatively distinct boundary detection is important for disease staging. In previous methods, bilinear interpolation has been recognized as a widely used technique for getting predictions with the same resolution as input images [11], [20], [29]. However, bilinear interpolation might not be suitable for nonlinear characters, especially non-rigid characters in pathological regions. At the end of our network, we have utilized the dense upsampling convolution [28] for boundary refinement.

Our contributions can be summarized as:

1). We have created a clinical gastric cancer segmentation dataset for our research, which has been delicately annotated by medical specialists.

2). We have proposed multi-scale embedding networks for segmenting cancerous regions of various sizes, in which we have integrated Atrous Spatial Pyramid Pooling module and encoder-decoder based semantic-level embedding networks.

3). We have applied the deformable convolutional module for adapting to the non-rigid characters of pathological images, and we have utilized the dense upsampling convolution for boundary refinement at the end of our architecture.

The rest of this paper is organized as follows. Section 2 presents a brief description of the related methods. Section 3 describes the dataset we have created. Section 4 introduces our proposed network architecture in detail. Section 5 describes the experiments and the ablation analyses. Section 6 is the conclusion of our work.

## II. RELATED WORK

### A. ADJUSTMENT OF RECEPTIVE FIELD

***Deep Neural Networks.*** Deep neural networks have been proved to be effective in object identification and semantic segmentation [29], [30]. Deep networks have more capabilities for feature representation. With repeated convolutional operations and pooling layers, neural networks would also obtain larger receptive fields. In recent years, nearly all segmentation tasks for digital pathology image analysis have chosen deep networks as baselines in their whole deep learning architectures. In this work, we have selected two widely employed models as our backbones, VGG19 [31] and ResNetV2 [22].

***Atrous Convolution.*** Atrous convolution has been firstly implemented in a dyadic wavelet transform method [32], which has been widely used in signal processing. In deep networks, the resolution of the final feature maps will be significantly reduced with the cascaded pooling layers and striding operations. This phenomenon is disadvantageous for our segmentation task. Motivated by obtaining a wider range of information under similar consumption, Yu and Koltun [17] have used atrous convolution to replace these resolution-reduced layers. They have proposed stacked atrous convolutional layers with increasing rates of dilation. In [19], a similar approach has been adopted for simultaneously increasing receptive fields and preserving the resolution of feature maps.

***Deformable Convolution.*** As we can see in the dataset images shown in Fig.1, cancerous regions are mostly non-rigid. These regions usually have fluidic shapes with irregular

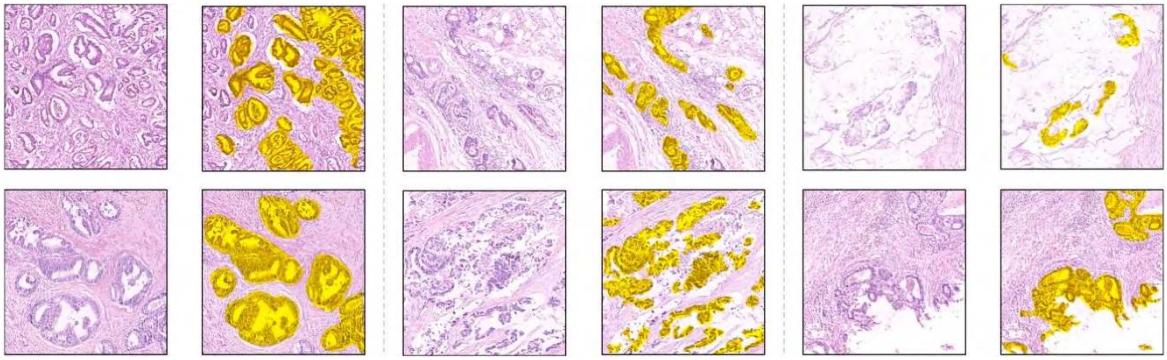


**FIGURE 1.** Six pairs of image-label samples from our gastric cancer segmentation dataset. Regions with yellow covered are cancerous. In these images, various sizes, vague boundaries, and non-rigid characters are shown clearly. Our dataset consists of 500 pathological images with delicate annotations.

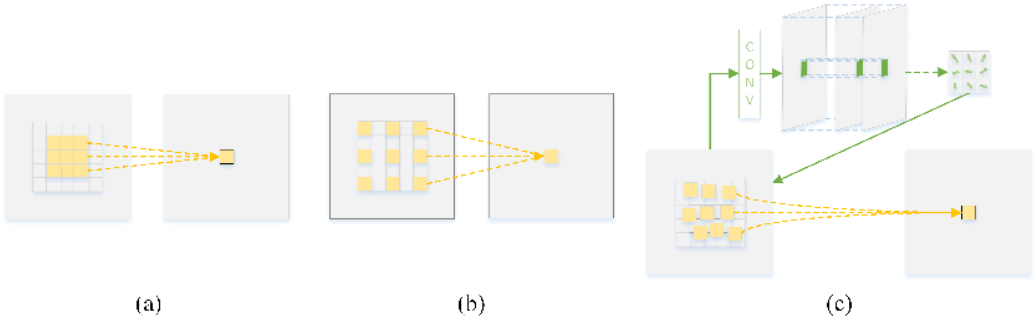



**FIGURE 2.** Three types of convolutional operations used in our network architecture. (a). Traditional convolution with kernel size = 3. (b). Atrous convolution with dilation = 1. The receptive field becomes nearly twice of ordinary convolution. (c). Deformable convolution. There are two branches in deformable convolution, the green branch is responsible for learning the offsets of the kernel coordinates. The yellow branch implements the proper convolutional operation.

boundaries, same as the regions of sky and lakes. Therefore, the traditional rectangle convolution may be limited for modeling unknown transformations or deformations, especially for boundaries [14]. The effective receptive field of the network would also decrease in this situation. In previous researches, super-pixel fusion [33] and region growing [34] have been proved as two efficient methods in pathology image segmentation or medical image analysis [35]–[37]. Meanwhile, these methods have been used to provide prior information in deep networks.

With the development of deep learning, the strategies of end-to-end learning and the thought of integration have become mainstream [38]. To address the limitation of rectangle convolution kernels, researchers have made some efforts on internal perception mechanism. Spatial transform networks [15] and deformable part models [16] are two advanced methods. In our approach, we have drawn on the experience of deformable convolution [14], which has integrated the deformable operation into convolutional layers. In Fig. 2, we could find that both the convolutional parameters and coordinate offsets could be learned in the networks, bringing in the adaptive receptive field.

### B. MULTI-SCALE EMBEDDING

***Semantic-level Embedding Networks.*** Image semantic segmentation is a synthetic task which demands contexture features ranging from global perception to detail-attention. Semantic-level embedding, or contexture feature fusion, has been employed in plenty of networks with different purposes and strategies. FCN [29] and deconvolution network [39] have employed deconvolution to restore the resolution of high-level features. SegNet [24] has used the indices of pooling during the encoder process, and U-Net [11] has employed skip connections from encoder to decoder layer-by-layer.

***Atrous Spatial Pyramid Pooling.*** Pyramid structure is a conventional technique in image processing [40]. With the development of deep learning, spatial pyramid pooling module has been firstly proposed in SPP-Net [41], aiming at detecting objects of various sizes. PSPNet [42] has used a similar strategy to serve the multi-scale feature extraction. Atrous Spatial Pyramid Pooling is an advanced structure based on atrous convolution and pyramid module. This module has been employed in DeepLab series [20], [43] to capture more comprehensive context information. Till this day, Atrous Spatial Pyramid Pooling has been one of the most

efficient modules in semantic segmentation for multi-scale embedding [43].

### C. DENSE UPSAMPLING CONVOLUTION

Coming from super-resolution tasks [44], [45], dense upsampling convolution is an approach of upsampling. In semantic segmentation tasks, bilinear interpolation and transposed convolution are two widely used techniques for final mask upsampling. Compared with these two, dense upsampling convolution has been proved to be a better replacement in fine-detailed information recovery [28]. In addition, with respect to the computational consumption, dense upsampling convolution is a tradeoff between bilinear interpolation and transposed convolution [28].

## III. DATASET

We have sampled our pathological images from clinical data. Under the 20 × optical magnification, we have obtained 500 pathological images with the resolution of 2048 × 2048. All the images have been cropped from whole pathological slides of gastric areas, with typical cancerous regions. The dataset has been delicately annotated by our cooperative medical specialists twice per image. Firstly, pixel-level cancerous segmentation masks have been provided by a group of experts. Then the annotated images have been finely refined by the other group, with the agreement of group one. Fig. 1 shows several examples of our dataset.

In our experiments, we have randomly divided the dataset into a training set and a testing set. The training set has been used to train the model parameters, and the testing set has been used to verify the effect of models. We have selected 350 images for training and 150 for testing. Before the images have been fed into the networks, we have cropped each 2048 × 2048 image of training set into four 1024 × 1024 patches. We have standardized the image-patches by subtracting the average pixel value of the dataset and dividing the standard deviation. In addition, basic image augmentation such as rotation, shifting, and affine transformation have been randomly used before each forward training procedure.

## IV. METHOD

In this section, we briefly introduce the architecture of the selected backbone [21] firstly. Then we describe the integrated encoder with atrous convolution and deformable convolution. Meanwhile, we review the encoder-decoder architecture and discuss the semantic-level embedding strategy. Finally, we recommend our proposed decoder module and the motivation of dense upsampling. In the end, we introduce our ultimate network and give a concrete illustration.

### A. SELECTED BACKBONE RESNET-101.

In this paper, we have selected a widely used deep network ResNet101 in our experiments. There are 5 stages with different output strides = {4, 8, 16, 32, 64} in the network, which is implemented mainly by 4 blocks. These 4 blocks are composed by several "bottleneck" resblocks [21] with different numbers of {3, 4, 23, 3}. For easier implementation, we have modified the places of downsampling convolutions to the last resblock in each block. We have selected the first four stages as the encoder in our segmentation network. In our experiments, we have paid more attention to the architecture of stage 3 and stage 4. For segmentation, we have used 16 as our output stride of the encoder. In stage 4, we have employed the dilated convolution with dilation = 2 (with kernel size > 1) for preserving the resolution.

**TABLE 1. Ablation analysis for pre-training.**

| Models | Pre-trained Dataset | Accuracy | mIOU |
|---|---|---|---|
| Mobile-FCN | None | 0.7857 | 0.6455 |
| Mobile-FCN | Imagenet | **0.8210** | **0.6952** |
| FCN-32s | None | 0.7891 | 0.6508 |
| FCN-32s | Imagenet | **0.8645** | **0.7614** |
| FCN-8s | None | 0.8215 | 0.7124 |
| FCN-8s | Imagenet | **0.8697** | **0.7695** |
| VGG-Unet | None | 0.8195 | 0.7145 |
| VGG-Unet | Imagenet | **0.8667** | **0.7648** |
| Res-Unet | None | 0.8354 | 0.7215 |
| Res-Unet | Imagenet | **0.8793** | **0.7786** |
| DeepLabV3 | None | 0.8430 | 0.7371 |
| DeepLabV3 | Imagenet | **0.8938** | **0.7986** |

**TABLE 2. Ablation analysis for deformable convolution.**

| Models | Output stride/Stage | Number of layers | Accuracy | mIOU |
|---|---|---|---|---|
| U-net (baseline) | - | - | 0.8667 | 0.7648 |
| Res-Unet (deform) | 4 / 1 | 3 | 0.8774 | 0.7757 |
| Res-Unet (deform) | 4 / 1 | 2 | 0.8755 | 0.7729 |
| Res-Unet (deform) | 8 / 2 | 3 | 0.8746 | 0.7662 |
| Res-Unet (deform) | 8 / 2 | 2 | 0.8752 | 0.7681 |
| Res-Unet (deform) | 16 / 3 | 3 | **0.8796** | **0.7782** |
| Res-Unet (deform) | 16 / 3 | 2 | 0.8756 | 0.7725 |
| DeepLabV3 (baseline) | - | - | 0.8938 | 0.7986 |
| DeepLabV3 (deform) | 16 / 3 | 3 | **0.8930** | **0.8016** |
| DeepLabV3+ (segment baseline) | - | - | 0.9030 | 0.8145 |
| DeepLabV3+ (deform) | 16 / 3 | 3 | **0.9079** | **0.8178** |

### B. ENCODER WITH DEFORMABLE CONVOLUTION AND ATROUS CONVOLUTION.

***Deformable convolution integration.*** The deformable convolution consists of two parts, ordinary convolution, and the deformable branch. The deformable branch achieves the main step for deformation, which learns offsets of coordinates in each convolutional kernel. In our final architecture, we have applied the deformable convolution at the last few resblocks in stage 3 (with kernel size > 1), as reported in Table. 2.

In this section, we utilize a 3 × 3 convolutional kernel to describe the deformable process. Consider a grid G in (1) on the previous feature maps X[i]. i is the coordinate ($i_h$, $i_w$) of

the feature maps. h and w represent the two dimensions of the feature maps. $G_i$ is the region convolved by the kernel. O is the $3 \times 3 \times 2$ offsets of nine coordinates in G on the specific location i, which is acquired by ordinary convolution operation in the deformable branch, as the green operations illustrated in Fig. 2(c).

$$\begin{aligned} G = \{&(-1,-1), \quad (-1,0), \quad (-1,1), \\ &(0,-1), \quad (0,0), \quad (0,1), \\ &(1,-1), \quad (1,0), \quad (1,1)\} \end{aligned} \tag{1}$$

$$\begin{aligned} I = \{&(i_h, i_w), \quad (i_h, i_w), \quad (i_h, i_w), \\ &(i_h, i_w), \quad (i_h, i_w), \quad (i_h, i_w) \\ &\times(i_h, i_w), \quad (i_h, i_w), \quad (i_h, i_w)\} \end{aligned} \tag{2}$$

$$G_i = G + I \tag{3}$$

$$New_i = G_i + O \tag{4}$$

Then the new locations of the nine pixels are $New_i$ for coordinate $G_i$. *Conv* is the function of ordinary convolution, and $k_i$ is the nine learnable weights of $3 \times 3$ kernels. We could obtain new values $X[New_i]$ of the new locations by bilinear interpolation on the overlapping pixels. Then we could get the value of the corresponding pixel $Y[i]$ in the final output $Y$ referring to (5). The illustration of deformable operation is shown in Fig. 2(c).

$$Y[i] = Conv(X[New_i], k_i) \tag{5}$$

In our experiments, we have first discussed the effectiveness of non-rigid feature perception on ResNet based U-net, to find a better-integrated strategy in the backbone. We have compared the applications on stage 1, 2, and 3 with output strides = 4, 8, and 16 respectively. To avoid memory workflow, we have selected 2 or 3 resblocks at the end of each block for deformable integration (with kernel size $> 1$). Then, based on the backbone with the best performance, we have applied the same integrated strategy to complicated frameworks in DeepLabV3 and DeepLabV3+.

In our final architecture, we have applied the deformable convolution on the penultimate block (block 3) of the encoder for output stride = 16.

***Atrous convolution integration.*** Atrous convolution, or dilated convolution, is a powerful and widely used operation in semantic segmentation. This operation can change the field of view explicitly while maintaining the effective receptive field of pre-trained deep classification framework to generate dense prediction mask.

Identical to the description of deformable convolution above, atrous convolution could also be defined as:

$$Y[i] = Conv(X[New_i], k_i) \tag{6}$$

In this situation, the new locations of the nine pixels are calculated by:

$$New_i = G_i + R \tag{7}$$

$$\begin{aligned} R = \{&(-r,-r), \quad (-r,o), \quad (-r,r), \\ &(0,-r), \quad (0,0), \quad (0,r), \\ &(r,-r), \quad (r,0), \quad (r,r)\} \end{aligned} \tag{8}$$

In Equation (8), $r$ is the dilation ratio of atrous convolution. In this way, we could find that though the motivations are different, atrous convolution is a special case of deformable convolution with fixed offsets as illustrated in Fig. 2(b). Meanwhile, we could comprehend atrous convolution as convolving the input $X$ with upsampled filters. These filters are constructed by inserting $r-1$ zeros between each pixel in the convolutional kernel.

In our architecture, we have applied this operation to take place of the traditional convolution (with kernel size = 3) in the last resblocks of block 3 and block 4 with dilation = 2. These two operations have been employed to remove the downsampling convolutional operation. In block 4, other convolutional operations with kernel size = 3 have also integrated the dilated strategy for preserving the receptive field.

***Atrous Spatial Pyramid Pooling.*** At the last of our used encoder module (after stage 4), there has been an efficient context module, called Atrous Spatial Pyramid Pooling module [20], [23], [39]. This module captures multi-scale information by applying global average pooling and atrous convolution with different dilation ratios parallelly, achieving impressive accuracy with considerable computational costs.

As the frameworks based on atrous convolution and ASPP module have proved their success, we have implemented an encoder with ASPP, using atrous convolution with different dilation ratios = {1, 6, 12, 18} parallelly.

### C. SEMANTIC-LEVEL EMBEDDING STRATEGY

Most encoder-decoder based frameworks [11], [26], [27] have used the structures shown in Fig. 3(b). Frequently, the encoder part is a deep network pre-trained on large-scale classification datasets such as Imagenet [46]. With feature maps of different semantic levels and resolutions, numerous feature fusion strategies have been proposed gradually. The decoder has fused these features to predict final segmentation masks. Generally, the feature fusion process could be formulated as:

$$y_l = F_l(x_l + Upsample(y_{l+1})) \tag{9}$$

where $y_l$ is the fused feature of $l-th$ level, and $x_l$ corresponds to the same semantic-level feature generated by the encoder. $F_l$ is a non-linear transformation for decoding the combined feature, which is usually a complex convolutional module.

In contrast, we have fused features of different semantic levels simultaneously and used a single-layer decoder to generate prediction results as illustrated in Fig 3(c). Different from the fusion strategy layer by layer abovementioned, we have first concatenated the feature maps from different semantic levels equally for sufficient utilization. Then the decoder has merged the feature maps together by a single-layer convolutional structure. More abstract and higher-level information has been extracted through the whole features. The effectiveness of this structure has been proved by DeeplabV3+ [43] in the field of semantic segmentation. In our tasks, this strategy has also achieved

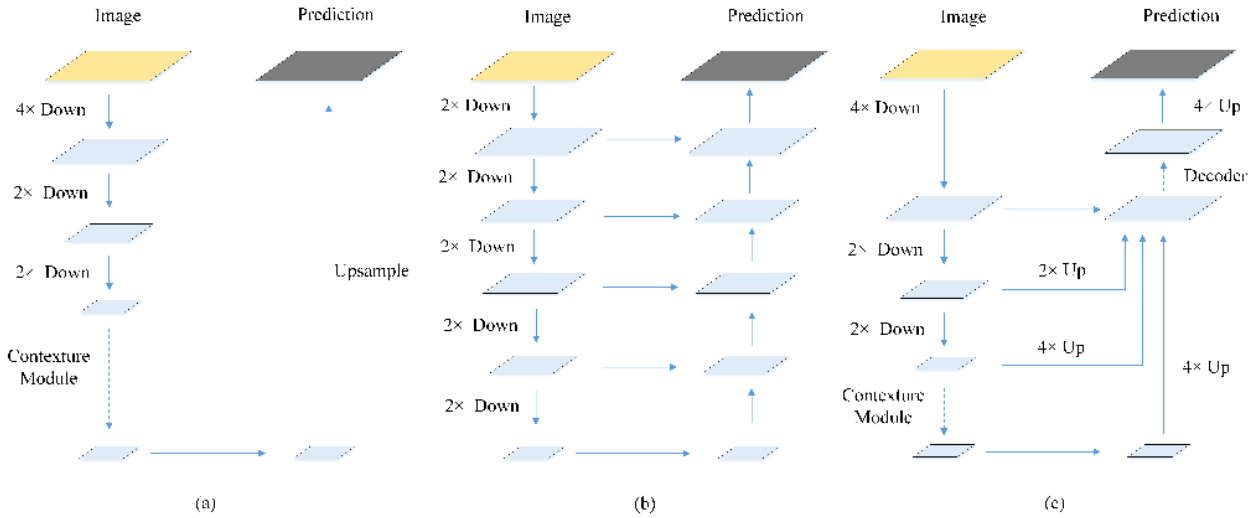



**FIGURE 3.** Different semantic-level embedding strategies. (a). Backbone based model without low-level feature fusion. (b). Encoder-decoder model with a complex stratified decoder. (c). Our proposed strategy, which fuses low-level and high-level features directly and then decodes those features with a lightweight decoder.

**TABLE 3.** Ablation analysis for semantic embedding.

| 1/4 feature | 1/8 feature | 1/16 feature (deform) | 1/16 feature (atrous) | Accuracy | mIOU |
|---|---|---|---|---|---|
| - | - | - | - | 0.8957 | 0.8046 |
| √ | - | - | - | **0.9079** | **0.8178** |
| √ | √ | - | - | 0.9056 | 0.8150 |
| √ | √ | √ | √ | 0.9089 | 0.8195 |
| √ | √ | √ | - | **0.9148** | **0.8210** |

better results compared with no-decoder architecture as shown in Table. 3. Furthermore, we have designed an advanced lightweight module for the decoder.

Our feature fusion strategy could be defined as:

$$y_l = F_l(x_l + Upsample(x_{l+1}) + Upsample(x_{l+2}) + \ldots) \quad (10)$$

In our networks, there are 3 different semantic levels of feature maps which could be extracted from stage1, 2 and 3, whose spatial resolutions are {1/4, 1/8, 1/16} of the input size. We have also extracted feature maps before and after the ASPP module, which are the same at the resolution but different at the semantic level. For reducing the channel number of the lower-level features, we have used $1 \times 1$ convolution to change the channel numbers to be the same as the final features after ASPP.

We have selected several subsets of these features to generate prediction results, and used them to retrain the whole system. In Section IV, we introduce these experiments in detail. We have found that combining the encoded features with output strides = 1/4, 1/8 and features after deformation with output strides = 1/16 would produce better performance.

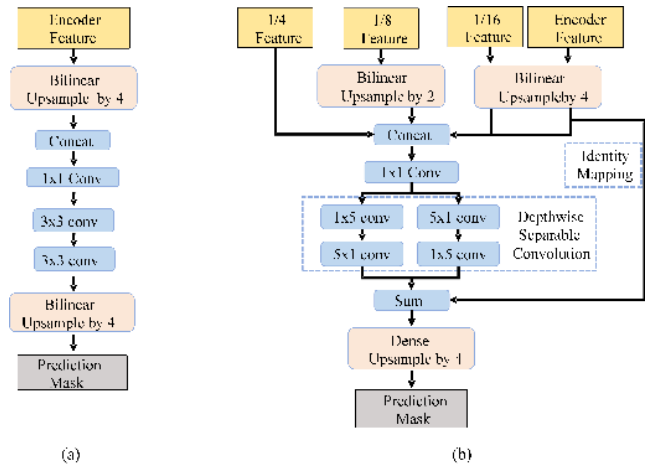



**FIGURE 4.** Illustration of the design of dense upsampling convolution at the end of the decoder. The feature maps are upsampled by 2.

### D. PROPOSED DECODER AND DENSE UPSAMPLING

***Proposed Decoder module.*** Over the previous methods, the predictions of segmentation masks computed from a one-layer decoder are usually obtained simply by bilinear upsampling [29]. In DeepLabV3+, a more complicated decoder has been proposed by adding some $3 \times 3$ convolutional layers, as shown in Fig 4(a). As a consequence, this model has achieved state-of-the-art performance in semantic segmentation under the same backbone [43]. For verifying the efficiency, we have tested this architecture on our gastric cancer segmentation dataset. As a result, this method has performed best in several existing advanced networks. We could see the results in Table. 4.

In this paper, we have replaced the $3 \times 3$ kernels with larger sizes $n \times n$ for gathering more information.

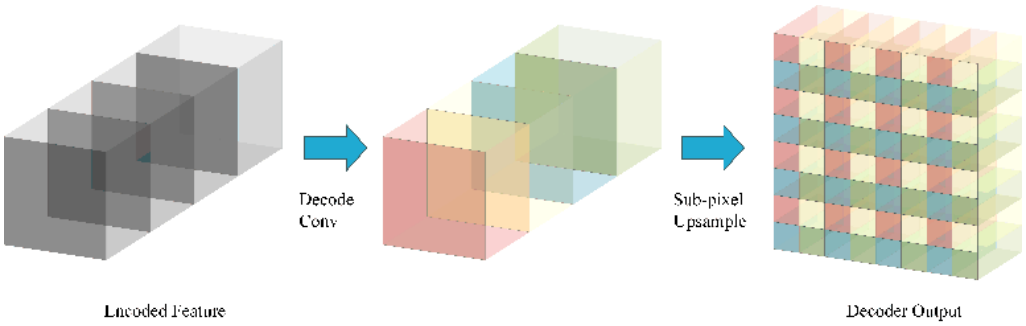



**FIGURE 5.** Decoder structures. (a) The baseline decoder with 2 layers of 3 × 3 convolutions in DeepLabV3+. (b) Our proposed decoder with two groups of separable convolution.

**TABLE 4.** Ablation analysis for decoder architecture.

| Architecture | Number | Identity mapping | Accuracy | mIOU |
|---|---|---|---|---|
| No-decoder | - | - | 0.8938 | 0.7986 |
| 3x3 | 2 | - | 0.9030 | 0.8145 |
| 1x7+7x1 | 1 | - | 0.8994 | 0.8069 |
| 1x7+7x1 | 2 | - | 0.9049 | 0.8150 |
| 1x5+5x1 | 1 | - | 0.9065 | 0.8170 |
| 1x5+5x1 | 2 | - | 0.9080 | 0.8184 |
| 1x5+5x1 | 2 | √ | **0.9103** | **0.8202** |

Considering the cost in computation and inspired by [47], we have used depthwise separable convolution to factorize the n × n convolution into a combination of 1 × n + n × 1 and n × 1 + 1 × n convolutions. In our experiments, we could find that this module has been both efficient and effective, as shown in Table. 3.

In addition, we have proposed a residual connection to refine boundary, since the robust and high-performance encoder can encode high-semantic and precise prediction. The restoration of boundary information can be modeled as a residual branch [26].

Our decoder is shown in Fig. 4(b).

***Dense Upsampling Convolution.*** Dense upsampling convolution module has been used in super-resolution [44], Low-light image processing [48] and semantic segmentation [28] for different purposes. The dense upsampling convolution could decode features at the original resolution by embedding upsampling operation in decoder channel groups. In the description of [28], the combination of dense upsampling and a single-layer convolution have shown superior capability as a replacement of bilinear interpolation, since it is capable of recovering fine-detailed information compared with common upsampling operations. Furthermore, the suitable employment of this operation would also make up for the gridding problem [28], which might be caused by atrous convolution.

Considering an input pathological image with resolution $H \times W$, the network would generate prediction logits with dimension $H \times W \times C$, where $C$ is the number of categories. In general, the feature map before making predictions usually has the resolution of $h \times w \times c$, where $h = H/s$, $w = W/s$, $s$ is the output stride, or downsample factor, of the feature map, and $c$ is the channel number usually greater than $C$. The core operation of dense upsampling convolution is to generate feature maps with the size of $h \times w \times (s \times s \times c)$, and then divide the whole feature map into equivalent $s \times s$ subsets across the channel. At last, the final $C$ prediction masks are

obtained by interleavedly reshaping of subsets with a softmax layer and an argmax layer following.

The operation of dense upsampling convolution is illustrated in Fig. 5. As we can see, the channel depth of the final output feature map is mapped to spatial grids.

In our network architecture, the decoder part designed for segmentation should be able to upsample feature intrinsically at the end of the network. Motivated by the above observations, we have designed our whole decoder as the combination of the proposed decoder module and dense upsampling operation. Compared with bilinear interpolation in [39], we have got more distinct boundaries shown in Fig. 7.

### E. OVERALL FRAMEWORK

Our overall model is shown in Fig. 6. We have used pre-trained ResNetV2 with atrous convolution, deformable convolution, and Atrous Spatial Pyramid Pooling module as our encoder. Semantic-level feature maps with different downsampling rates have been extracted from different stages of the encoder network. The lightweight decoder has fused the multi-scale features and provided the low-resolution score maps. Then we have used the dense upsampling convolution to upsample the score maps for dense prediction. The details can be referred to the illustration in Fig. 6.

## V. EXPERIMENTS

### A. IMPLEMENTATION DETAILS

Our proposed model has been built on Tensorflow [49] framework. The backbone network has been the ResNetV2 101 model pre-trained on Imagenet. For adaptation to our architecture, we have resized the input images to 512 x 512. We have followed common dataset augmentation strategies to train our model. We have randomly cropped each resized image with scales in the interval [0.5, 1.0]. The pre-trained

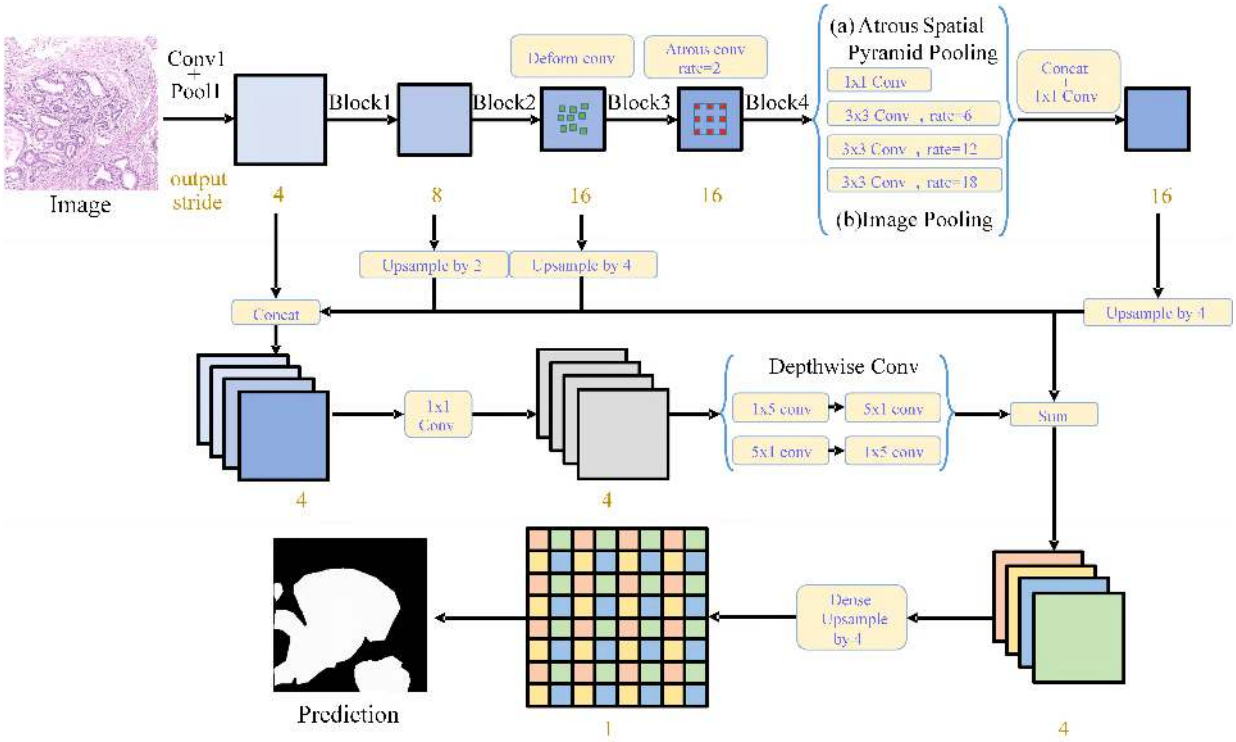



**FIGURE 6.** Overall framework of our approach. We have illustrated the whole architecture in three stages. First stage (above the first light blue line): The ResNet based encoder integrated with specific convolution and multi-scale ASPP module. Second stage (between the two light blue lines): The semantic embedding architecture and the lightweight decoder. Third stage (under the second light blue line): The final dense upsampling operation.

context convolution operations all have included batch normalization.

Though greater batch size has been proved to be effective in semantic segmentation, we have chosen a fixed 8 batch size for searching more efficient architectures. The limitation of the power of computation has also been considered simultaneously. For back propagation, we have selected the cross-entropy loss function at each pixel over the categories and used standard Stochastic Gradient Descent (SGD) with weight decay 2e-4. Inspired by DeepLab [18], we have used the ''poly'' learning rate policy. The learning rate has changed along with the training steps by multiplying:

$$(1 - \frac{iter}{maxiter})^{power} \tag{11}$$

in which *iter* represents the training steps and *maxiter* represents the whole training iterations. we have set the $power = 0.9$ and initial learning rate 1e-3. All the matrix calculations have been implemented on 2 NVIDIA TITAN X GPUs. The rest of the calculations have been processed by 1 Intel Core I7-6900k CPU with octa-core and 3.2GHz clock speed. The size of the calculated memory is 64G.

### B. EVALUATION INDICATORS

Our performance has been measured by accuracy and mean intersection-over-union across the classes of cancerous regions and normal regions.

Accuracy is the ratio of the number of correctly predicted pixels and the total number of pixels. Mean Intersection-over-Union is a standard indicator for semantic segmentation, it calculates the ratio of the intersection and union of the real values (ground truth) and the predicted values (predictive segmentation).

### C. ABLATION ANALYSIS

In the field of semantic segmentation, enough computing power with larger batch size would be an effective way to get better performance. In our experiments, we have run a number of ablations to analyze our network under the same computational resources and 8 batch size for finding the most efficient architecture.

We have discussed the thought of transfer learning by using the models pre-trained on natural image dataset [50]. We have analyzed the effect of the deformable module and atrous-based architectures. Also, we have discussed the strategies of feature fusion. Then we have compared our proposed decoder and the upsampling process with previous architecture. Finally, quantitative comparisons against several prior advanced methods have demonstrated the superiority of our approach.

***Pre-trained models.*** Our segmentation task on pathological images is distinctly different from conventional semantic segmentation tasks on natural images, though they could use the same architecture. In previous studies, using pre-trained

models on semantic segmentation subtasks has become the mainstream, but there are still doubts whether pre-trained models on Imagenet are beneficial for pathological segmentation [50], since the wide difference between data distributions of these two datasets.

We have compared several progressive methods on our dataset with and without pre-training. In conclusion, we have found that with pre-trained models, all performances of the methods have approximately increased at least by 4%. The pre-training strategy is beneficial for our pathological task. Detailed quantitative information is shown in Table. 1.

***Deformable module.*** We have used the deformable module to enhance the capability of learning pathological non-rigid features. Referring to [14] and considering the limitation of computational resources, we have applied this convolution on limited layers at last of each block (with kernel size $= 3$).

For verifying the effectiveness of deformable convolution, we have implemented two groups of experiments on ResNet based U-net and our primitive ResNet based baselines DeepLabV3 and DeepLabV3+.

In the experiments on U-net, we have experimented on each block of the encoder with the same number of layers in stage 1, 2 and 3. We have selected 2 or 3 resblocks at the end of each block for deformable integration (with kernel size $> 1$). The results are shown in Table. 2. We could find that the deformation with larger output stride/stage and more layers have performed better. Therefore, we have chosen the deformation integrated encoder with the best performance to replace the plain encoder in DeepLabV3 and DeepLabV3+.

We have compared these two models with and without deformable convolution. We could find that deformable convolution has also shown better capability for pathological feature learning. In our final architecture, we have integrated the deformable module on the penultimate block for output stride $= 16$, with the last 3 layers (with kernel size $= 3$) in this block.

***Semantic-level embedding strategies.*** We have extracted the low semantic-level but high-resolution features with downsample factor of 1/4 and 1/8 at last of block1 and block2 respectively. We have also extracted features of the final layer in block3 with downsample factor of 1/16 and used the features before the Atrous Spatial Pyramid Pooling module (after the atrous convolution), which have the same resolution with downsample factor of 1/16 but different semantic level. The feature map before ASPP is coarser and without multi-scale information. For reducing the channel number of the

lower-level features, we have used $1 \times 1$ convolution to change the channel numbers to be the same as the final features after ASPP.

All these feature fusion strategies have concatenated subsets of the features abovementioned with the feature maps after the multi-scale Atrous Spatial Pyramid Pooling module. We have selected several subsets of these features to generate prediction results, and used them to retrain the whole system. We have compared these different fusion structures in experiments. In Table. 3, we could clearly see that fusing low semantic-level but high-resolution features with multi-scale and high-semantic encoded feature could significantly improve the performance. Finally, we have chosen the feature fusion strategy of best performance. Also, it is a remarkable fact that the second model fusing encoded features with output strides $= 1/4$ is the deformable DeepLabV3+ in Table. 2.

**TABLE 5.** Ablation analysis for dense upsampling convolution.

| Models | Upsampling Methods | Accuracy | mIOU |
|---|---|---|---|
| DeepLabV3 | Bilinear | 0.8938 | 0.7986 |
| DeepLabV3 | Dense | 0.8920 | 0.7976 |
| DeepLabV3+ | Bilinear | 0.9030 | 0.8145 |
| DeepLabV3+ | Dense | 0.9069 | 0.8150 |
| Proposed | Bilinear | 0.9148 | 0.8210 |
| Proposed | Dense | **0.9160** | **0.8265** |

**TABLE 6.** Comparison wth previous methods.

| Models | Accuracy | mIOU |
|---|---|---|
| Mobile-FCN | 0.8210 | 0.6952 |
| FCN-32s | 0.8645 | 0.7614 |
| FCN-8s | 0.8697 | 0.7695 |
| VGG-Unet | 0.8667 | 0.7648 |
| Res-Unet | 0.8793 | 0.7786 |
| DeepLab | 0.8770 | 0.7758 |
| DRN | 0.8810 | 0.7812 |
| DeepLab+ASPP | 0.8819 | 0.7930 |
| DeepLabV3 | 0.8938 | 0.7986 |
| DeepLabV3+ | 0.9030 | 0.8145 |
| Proposed | **0.9160** | **0.8265** |

***Proposed decoder module.*** In this subsection, we have compared our proposed decoder module shown in Fig. 4(b) with the previous network DeepLabV3+ as shown in Fig.4(a) and the no-decoder network DeepLabV3, using the same bilinear interpolation at the end.

We have examined the effect of different decoder architectures with different types of separable convolution and tested the connection of identity mapping.From Table. 3, we could see that parallelly repeating $1 \times 5 + 5 \times 1$ convolution twice would generate the best predictive masks. Also, the identity mapping, or residual refinement connection, has improved the performance as well. The performances are shown in Table. 4

Our final proposed decoder is the combination of the parallelly repeated $1 \times 5 + 5 \times 1$ convolutions and residual refinement connection.

***Dense upsampling.*** For refinement of details, we have replaced the bilinear interpolation with dense upsampling convolution, as a result, we have got approximate performance but better boundaries, as illustrated in Fig. 7. The proposed method is the architecture abovementioned which consists of deformable convolution, dilated convolution, ASPP module, and the semantic embedding strategy.

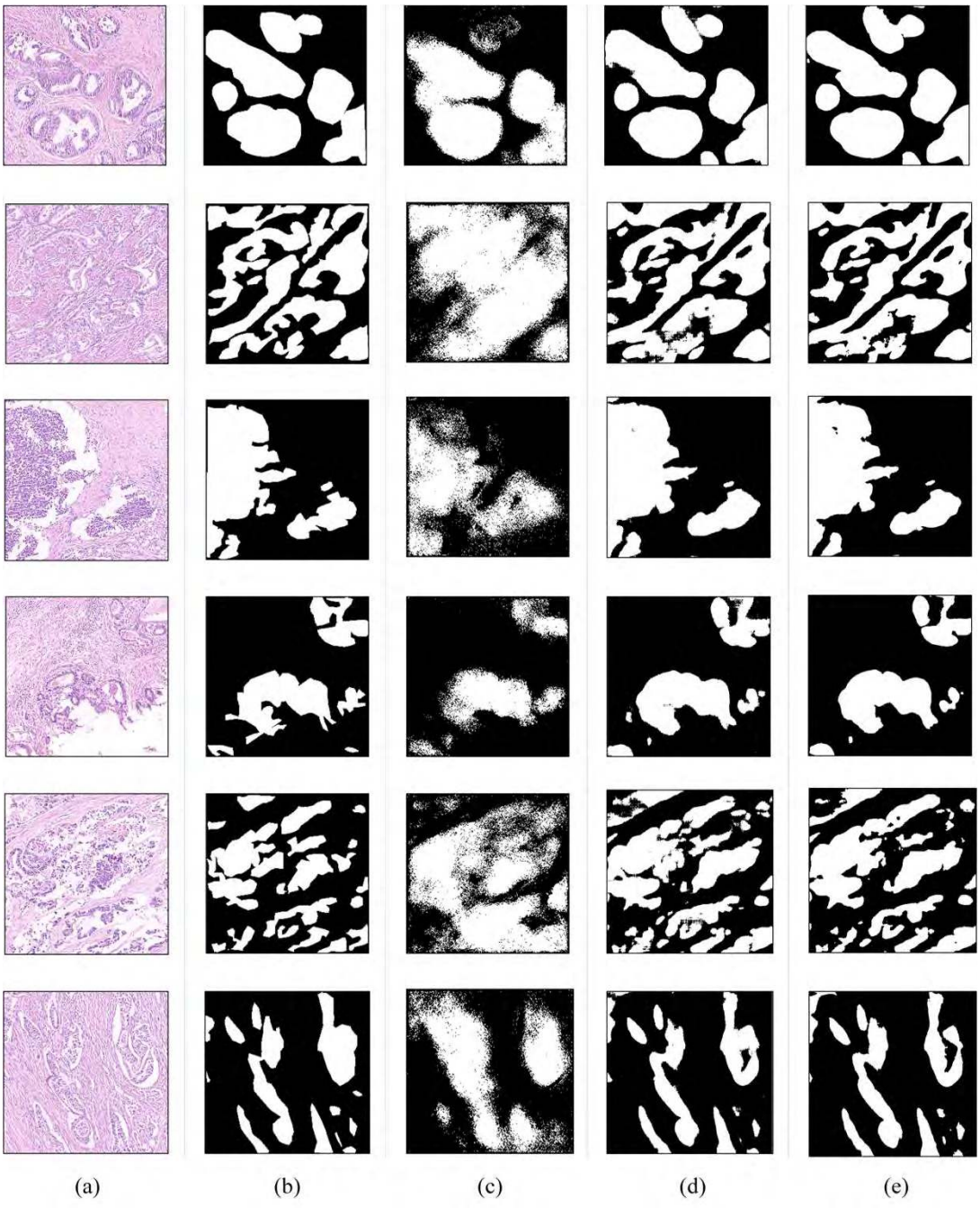



**FIGURE 7.** Visualization and comparison of several representative results on our test dataset. (a) Pathological images. (b) Ground truths of the images. (c) Predictions of ResNet based U-net. (d) Predictions of DeepLabV3+ with bilinear upsampling. (e) Our proposed method with dense upsampling.

The performances have been shown in Table. 5. We could find that integrating the dense upsample convolution at the last of our proposed architecture has shown the superior property.

***Comparison with previous methods.*** Our ultimate proposed method has been shown in Fig.6. The comparisons with previous advanced methods have been shown in Table. 6. We could find that our method has shown the best performance.

### D. RESULTS AND DISCUSSION

***Results.*** As shown in Fig. 7, we visualize some representative segmentation results of our proposed model and other related models. It is clear that our method has distinct advantages over the contrastive models. Our method is more

sensitive to regions with different sizes and handles the fine-grained features more carefully.

***Limitations.*** Though our method has got satisfactory predictions, there are still some limitations. Firstly, our dataset

is just enough for this research. All the images have been sampled under the same optical magnification and staining method. In clinical pathological image analysis, the diagnosis will face more complicated situations. There is still a lot of comprehensive work we need to do from experiments to clinical. Second, due to our limited computational budget, we have set the batch size = 8, which may not take full advantage of batch normalization. We believe that our performance will be better if the hardware is more powerful.

## VI. CONCLUSION

In this paper, we have proposed a deep learning architecture for gastric cancer segmentation which demonstrates the advantage of jointly utilizing multi-scale modules and specific convolutional operations. The well-defined framework can simultaneously learn the representations of regions with different sizes and non-rigid characters. Extensive comparative evaluations on our own dataset demonstrate that the proposed method is more accurate and efficient. Meanwhile, we need to do more both on datasets and algorithms for promoting the fusion of deep learning technology and pathological diagnosis.

## REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *Ca-Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, Nov./Dec. 2018.

[2] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1550–1560, Jul. 2017.

[3] D. Baltissen, T. Wollmann, M. Gunkel, I. Chung, H. Erfle, K. Rippe, and K. Rohr, "Comparison of segmentation methods for tissue microscopy images of glioblastoma cells," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Washington, DC, USA, Apr. 2018, pp. 396–399.

[4] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2016, pp. 460–468.

[5] L. Bi, J. Kim, A. Kumar, M. Fulham, and D. Feng, "Stacked fully convolutional networks with multi-channel learning: Application to medical image segmentation," *Vis Comput*, vol. 33, nos. 6–8, pp. 1061–1071, Jun. 2017.

[6] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2017, pp. 408–416.

[7] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2017, pp. 399–407.

[8] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016, *arXiv:1606.05718*. [Online]. Available: https://arxiv.org/abs/1606.05718

[9] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, Andre, L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, no. 10, pp. 1559–1567, Sep. 2018.

[10] Y. Li, X. Li, X. Xie, and L. Shen, "Deep learning based gastric cancer identification," in *Proc. Int. Symp. Biomed. Imag.*, Washington, DC, USA, Apr. 2018, pp. 182–185.

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

[12] S. Mejbri, C. Franchet, I. A. Reshma, J. Mothe, P. Brousset, and E. Faure, "Deep analysis of cnn settings for new cancer whole-slide histological images segmentation: The case of small training sets," in *Proc. 6th Int. Conf. Bioimag.*, Feb. 2019, pp. 120–128.

[13] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.

[14] J. F. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 764–773.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2017–2025.

[16] Y. Jeon and J. Kim, "Active convolution: Learning the shape of convolution for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1846–1854.

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent*, May 2016, pp. 1–13

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," in *Proc. Int. Conf. Learn. Represent*, 2015, pp. 1–14.

[19] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 636–644.

[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016 pp. 770–778.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 630–645.

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

[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017. doi: 10.1109/TPAMI.2016.2644615.

[25] G. Lin, A. Milan, C. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5168–5177.

[26] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1743–1751.

[27] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2016, pp. 519–534.

[28] P. Q. Wang, P. F. Chen, Y. Yuan, D. Liu, Z. H. Huang, X. D. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl Comput. Vis.*, Mar. 2018, pp. 1451–1460.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," Feb. 2014, *arXiv:1312.6229*. [Online]. Available: https://arxiv.org/abs/1312.6229

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent*, Sep. 2014, pp. 1–14.

[32] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Proc. Wavelets*, 1990, pp. 286–297.

[33] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[34] H. Nazeran, F. Rice, W. Moran, and J. Skinner, "Biomedical image processing in pathology: A review," *Australas Phys. Eng. Sci. Med.*, vol. 18, no. 1, pp. 26–38, Mar. 1995.

[35] W. Qin, J. Wu, F. Han, Y. Yuan, W. Zhao, B. Ibragimov, J. Gu, and L. Xing, "Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation," *Phys. Med. Biol.*, vol. 63, no. 9, p. 0950, May 2018.

[36] N. S. M. Raja, S. L. Fernandes, N. Dey, S. C. Satapathy, and V. Rajinikanth, "Contrast enhanced medical MRI evaluation using Tsallis entropy and region growing segmentation," in *Proc. J. Ambient Intell. Humanized Comput.*, May 2018, pp. 1–12.

[37] S. Ji, B. Wei, Z. Yu, G. Yang, and Y. Yin, "A new multistage medical segmentation method based on superpixel and fuzzy clustering," *Comput. Math. Methods Med.*, vol. 2014, Mar. 2014, Art. no. 747549.

[38] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Automat. Comput.*, vol. 14, no. 2, pp. 119–135, Apr. 2017.

[39] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, May 2015, pp. 1520–1528.

[40] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, Nov. 1984.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015. doi: 10.1109/TPAMI.2015.2389824.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*. [Online]. Available: https://arxiv.org/abs/1802.02611

[44] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[45] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," 2017, *arXiv:1707.02937*. [Online]. Available: https://arxiv.org/abs/1707.02937

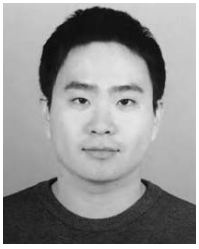
[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[48] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3291–3300.

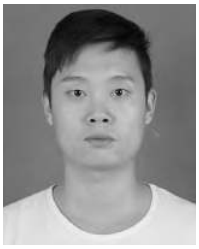
[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement.*, Savannah, GA, USA, vol. 16, Nov. 2016, pp. 265–283.

[50] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016. doi: 10.1109/TMI.2016.2528162.

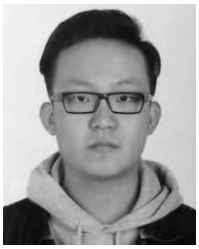


**MUYI SUN** received the B.S. degree from the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree in control science and engineering with the School of Automation. His research interests include pattern recognition, pathological image analysis, and computer vision.

**GUANHONG ZHANG** is currently pursuing the master's degree in pattern recognition and intelligent systems from the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include machine learning, computer vision, neural networks, and so on.



**HAO DANG** received the M.S. degree in pattern recognition and intelligent systems from the Henan University of Technology, Zhengzhou, China, in 2016. He is currently pursuing the Ph.D. degree in control science and engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include the pattern recognition, intelligent systems, machine learning, and so on.



**XINGQUN QI** is currently pursuing the master's degree in control science and engineering from the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include deep learning, computer vision, and so on.

**XIAOGUANG ZHOU** received the M.S. degree from the Department of Precision Instrument, Tsinghua University, in 1984, and the Ph.D. degree in engineering from the Tokyo University of Agriculture and Technology, Japan. He was a Visitor Professor with the Tokyo University of Agriculture and Technology, from 2001 to 2002, and a JSPS Researcher with Tokyo University, from 2013 to 2014. He is currently a Professor, and a Doctoral Supervisor with the School of Automation, Beijing University of Posts and Telecommunications. He also serves as the Director of the Engineering Research Center of Information Networks, Ministry of Education. He is the author of over 10 books, over 100 articles, and over 16 inventions. His research interests include control theory and its application in engineering, the Internet of Things and automated logistics systems, and mechatronics technology. He is a permanent member of the Chinese Association of Automation/Manufacturing Technology Committee and the China Institute of Communications/Equipment Manufacturing Technical Committee.

**QING CHANG** received the M.D. degree from the Department of Clinical Medicine, Nantong Medical College, in 2000, and the Ph.D. degree in molecular cell biology from the University of Tokyo, in 2013. He was a Postdoctoral Fellow with the Shanghai Jiaotong University School of Medicine, from 2013 to 2015. He is currently an Associate Professor, and a Master Supervisor with the Shanghai University of Medicine and Health Sciences. He also serves as an Attending Doctor and a Research Fellow of the Shanghai General Practice Medical Education and Research Center, Jiading District Central Hospital Affiliated Shanghai University of Medicine and Health Sciences. He has published over 20 articles. His research interests include the usage and challenge of innovating technology in general practice medicine, such as sequencing, big-data, and AI.