Region-aware Global Context Modeling for Automatic Nerve Segmentation from Ultrasound Images

Huisi Wu[1]*, Jiasheng Liu[1], Wei Wang[1], Zhenkun Wen[1], Jing Qin[2]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract

We present a novel deep learning model equipped with a new region-aware global context modeling technique for automatic nerve segmentation from ultrasound images, which is a challenging task due to (1) the large variation and blurred boundaries of targets, (2) the large amount of speckle noise in ultrasound images, and (3) the inherent real-time requirement of this task. It is essential to efficiently capture long-range dependencies by global context modeling for a segmentation network to overcome these challenges. Traditional global context modeling techniques usually explore pixel-aware correlations to establish long-range dependencies, which are usually computation-intensive and greatly degrade time performance. In addition, in this application, pixel-aware modeling may inevitably introduce much speckle noise in the computation and potentially degrade segmentation performance. In this paper, we propose a novel region-aware modeling technique to establish long-range dependencies based on different regions to improve segmentation accuracy while maintaining real-time performance; we call it region-aware pyramid aggregation (RPA) module. In order to adaptively divide the feature maps into a set of semantic-independent regions, we develop an attention mechanism and integrate it into the spatial pyramid network to evaluate the semantic similarity of different regions. We further develop an adaptive pyramid fusion (APF) module to dynamically fuse the multi-level features generated from the decoder to refining the segmentation results. We conducted extensive experiments on a famous public ultrasound nerve image segmentation dataset. Experimental results demonstrate that our method consistently outperforms our rivals in terms of segmentation accuracy. The code is available at https://github.com/jsonliu-szu/RAGCM.

Introduction

Surgery not only brings discomfort to the patient but also often causes severe post-surgical pain. Currently, the most commonly undertaken solution in clinical practice is to use anesthetics to relieve patient pain, which, however, may bring a bevy of undesirable side effects. Accurate nerve segmentation based on ultrasound images is a key step in inserting an indwelling catheter, which can greatly reduce dependence on anesthetics and speed up patient recovery (Baby and Jereesh 2017). Traditional ultrasound nerve segmentation relies on manual annotation by experienced doctors, which is tedious, time-consuming and subjective. To the end, automatic segmentation approaches are highly demanded in clinical practice. However, automatic nerve segmentation remains a very challenging task due to (1) the large variation and blurred boundaries of targets, (2) the large amount of speckle noise in ultrasound images, and (3) the inherent real-time requirement of this task.

In last decade, despite convolutional neural networks (CNNs) have achieve remarkable success in medical image segmentation, most of them are limited by local receptive fields, which are incapable of tackling large variation and blurred boundaries of targeting objects. To extract richer contextual information, several approaches employ multi-scale context fusion to capture long-range dependencies have been proposed, such as large-size filters (Szegedy et al. 2015), dilated convolution (Chen et al. 2018), pyramid pooling (Zhao et al. 2017), etc. However, recent studies (Luo et al. 2016) show that the effective receptive field of most of these networks is much smaller than theoretical. Recently, attention mechanism (Vaswani et al. 2017) has been widely investigated to capture long-range dependencies and many networks based on it have been proposed (Wang et al. 2018; Fu et al. 2019) to improve segmentation accuracy. However, most of these networks employ pixel-wise modeling to establish long-range correlation and use all positions to figure out the attention map (Li et al. 2019; Huang et al. 2019), which usually has high computation complexity and occupies a huge amount of memory in GPUs. In addition, studies reported in (Cao et al. 2019) show that, on the one hand, for different query positions in an image, the global context modeled by the non-local networks is almost the same. On the other hand, more redundant features will inevitably introduce more irrelevant noisy interference to degrade the segmentation performance, which is in particular severe for ultrasound images full of speckle noise.

In this paper, we present a novel deep learning model equipped with a new region-aware global context modeling technique in order to improve the segmentation accuracy while maintaining real-time performance. The proposed region-aware modeling technique is able to establish long-range dependencies based on a set of semantic regions instead of every pixel, and hence greatly reduce computation-

*Corresponding author. Email: hswu@szu.edu.cn.

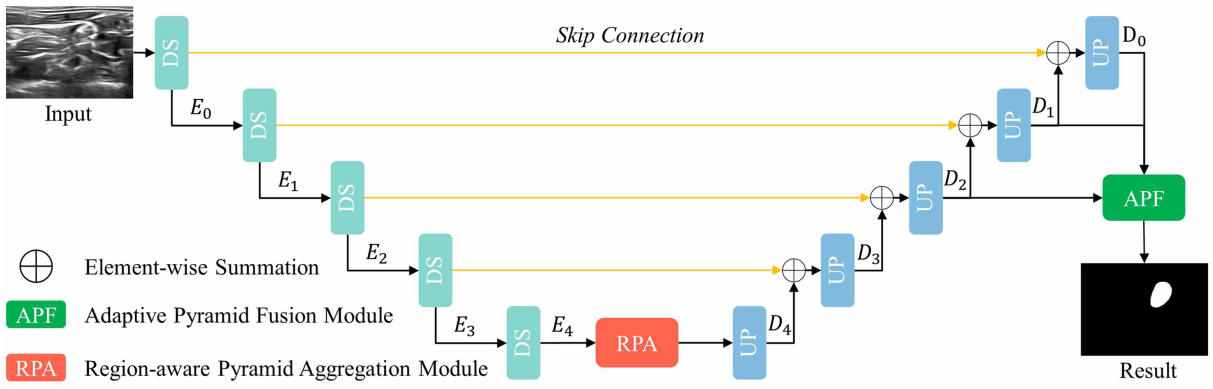

Figure 1: The overall architecture. RPA module can effectively distinguish the target regions from the blur background areas based on a novel region-level attention mechanism. APF module is used to dynamically fuse the multi-level feature maps from different layers and adaptively optimize the segmentation result.

al costs; we call it region-aware pyramid aggregation (RPA) module. In order to adaptively divide the feature maps into a set of regions, we develop an attention mechanism and integrate it into the spatial pyramid network to evaluate the semantic similarity of different regions. We further develop an adaptive pyramid fusion (APF) module to dynamically fuse the multi-level features to refining the segmentation results. We conducted extensive experiments on a famous public ultrasound nerve image segmentation dataset and experimental results demonstrate that our method achieves much better performance than our state-of-the-art rivals. Our main contributions can be summarized as follows:

- We propose a novel network for automatic nerve segmentation from ultrasound images and set state-of-the-art performance on a famous dataset.
- We propose a new region-aware global context modeling technique to establish long-range dependencies without compromising time performance and introducing much background noise in the fused feature maps; the proposed technique is general enough to be used to handle noisy images, such as ultrasound images, in real-time applications, such as many ultrasound-guided interventions.

Related Work

Nerve segmentation from Ultrasound Images A lot of effort has been dedicated to addressing the challenges of nerve segmentation from ultrasound images. In an early study, Hadjerci et al. (Hadjerci et al. 2016) employed a machine learning-based framework via hand-crafted features to handle this task. Kakade et al. (Kakade and Dumbali 2018) proposed a linear gabor binary pattern to pre-process ultrasound images and fed them into an ANN for detection and segmentation. Recently, CNN-based methods have been developed to improve segmentation accuracy. For example, Liu et al. (Liu et al. 2018) proposed a segmentation network based on a well-established DNN and employed a discriminator network to assess the segmentation quality to guide the segmentation network towards a better performance. However, most of these previous works do not take global contextual information into consideration to improve discrimination capability and exclude background noises.

Global Context Modeling It is evident that global multi-scale context modeling is beneficial to improve segmentation accuracy of deep networks. ParseNet (Liu, Rabinovich, and Berg 2015) first employed the global average pooling (GAP) operation to augment the features at each location for semantic segmentation. To exploit the global contextual information, PSPNet (Zhao et al. 2017) further extended it to the spatial pyramid pooling. For the same purpose, Deeplab (Chen et al. 2016) proposed an atrous spatial pyramid pooling module to capture the contextual information at multiple scales. After that, by combining the residual multi-kernel pooling (RMP) block and the dense atrous convolution (DAC) block, CE-Net (Gu et al. 2019) further captured more advanced semantic information while retaining spatial information. However, the local receptive field of traditional CNNs limited them from capturing more powerful correlations in a global perspective. Also, NLNet (Wang et al. 2018) also proposed to tackle this problem by pixel-wise modeling, but these approaches are time-consuming and computation-intensive, and hence not applicable in real-time applications. In addition, pixel-wise modeling techniques (Fu et al. 2019; Hou et al. 2020) may inevitably introduce more background noise in feature maps, making them not suitable for ultrasound image processing.

Method

Network Architecture

The Network architectureis illustrated in Figure 1, where two novel components are equipped to the well-validated encoder-decoder architecture (Ronneberger, Fischer, and Brox 2015), including RPA module and APF module. To achieve a compact yet precise structure, we adopt a pre-trained ResNet-34 (He et al. 2016) network as backbone, where both average pooling layer and the fully connected layer are removed. To capture multi-scale contextual information, we propose an RPA module by utilizing the self-attention mechanism to encode the region-wise features

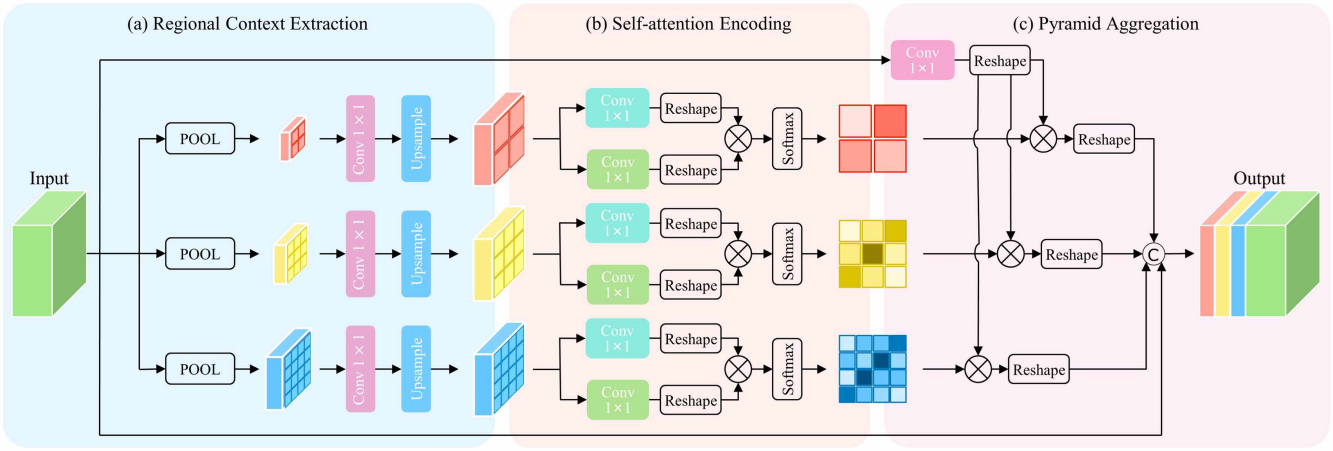

Figure 2: Our RPA module. Given the input feature map, we first generate three different regional feature maps based on three different-size pooling kernels. Self-attention encoding is then used to establish long-range dependence for each regional feature map. After pyramid aggregation of multiple branch regional features, we can obtain the output feature, which enhances the feature discrimination with regional-based long-range dependencies.

based on the pyramid pool sampling strategy, which also obtains more discriminative regional features. On the other hand, to achieve more effective multi-level pyramid feature fusion, we further introduce an APF module to dynamically select and fuse the dominant output features based on an efficiency channel attention mechanism. Finally, more refined segmentation results can be generated based on the more discriminative features extracted from the APF module.

Region-aware Pyramid Aggregation Module

To overcome the challenges of scale variations and a large amount of speckle noise, we need effectively capture long-range dependencies by global context modeling. Unfortunately, existing global context modeling techniques usually explore pixel-aware correlations to establish long-range dependencies, which are not only computation-intensive but also inevitably introducing more speckle noises to potentially degrade the segmentation accuracy. In this paper, inspired by the self-attention mechanism (Vaswani et al. 2017), we propose a region-aware pyramid aggregation (RPA) module to establish the long-distance dependence via a more effective regional contextual modeling. Specifically, our RPA module is implemented as following three steps.

Regional Context Extraction Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ extracted in the top layer of our encoder, we first employ three adaptive average pooling with different kernel sizes (3×3, 7×7, and 11×11,) to generate three different regional feature maps $\mathbf{X}_1 \in \mathbb{R}^{C\times3\times3}$, $\mathbf{X}_2 \in \mathbb{R}^{C\times7\times7}$, and $\mathbf{X}_3 \in \mathbb{R}^{C\times11\times11}$ respectively, where H, W, and C are the height, width, and the channel number of the feature map. To align above three regional feature maps to a uniform dimension, three different transformation functions followed by a upsampling operation are learned to restore $\mathbf{X}_i$ to the same spatial resolution as the original input feature map, which can be written as

$$\mathbf{X}'_i = \mathbf{Up}(\psi(\mathbf{X}_i, \theta_i)) \tag{1}$$

where $\mathbf{X}'_i \in \mathbb{R}^{\frac{C}{r}\times H\times W}, i = 1, 2, 3$. $\mathbf{Up}(\cdot)$ denotes the bilinear upsampling. $\psi(\cdot)$ is a projection function implemented by a 1×1 convolution followed by a ReLU activation function. θ_i is the relevant learnable parameter of $\psi(\cdot)$. r is the dimension reduction ratio. In our experiments, we set r to 4.

Self-attention Encoding To further enhance the regional feature representation, we also introduce a self-attention mechanism to establish long-range dependence for each regional feature map. Firstly, we utilize two 1×1 convolutions to squeeze the regional feature maps respectively. To satisfy the following self-attention multiplication, the dimensions of squeezed feature maps are also reshaped as following,

$$\mathbf{\Gamma}_i = \mathbf{Reshape}(\xi(\mathbf{X}'_i, \vartheta_i)), \mathbf{\Phi}_i = \mathbf{Reshape}(\varphi(\mathbf{X}'_i, \mu_i)) \tag{2}$$

where $\xi(\cdot)$ and $\varphi(\cdot)$ are two 1×1 convolutions. $\{\mathbf{\Gamma}_i, \mathbf{\Phi}_i | i = 1, 2, 3\} \in \mathbb{R}^{\frac{C}{4}\times(H\times W)}$. ϑ_i and μ_i are the related parameters. Therefore, we can perform a self-attention matrix multiplication operation on $\mathbf{\Gamma}_i$ and $\mathbf{\Phi}_i$ to obtain a regional attention map $\mathbf{\Pi}_i$, written as

$$\mathbf{\Pi}_i = \mathbf{\Gamma}_i^{\mathrm{T}} \times \mathbf{\Phi}_i \tag{3}$$

where $\{\mathbf{\Pi}_i | i = 1, 2, 3\} \in \mathbb{R}^{(H\times W)\times(H\times W)}$. We also use a softmax operation to produce the corresponding weighting map for each regional feature, which can be expressed by the following,

$$\mathbf{\Omega}_{(i,j)} = \frac{e^{Z_{(i,j)}}}{\sum_{j=1}^{k} e^{Z_{(i,j)}}} \tag{4}$$

where j is one pixel of the i^{th} regional feature map, and k is the number of channels in the regional attention map. Finally, based on a matrix multiplication between the $\mathbf{X}$ and $\mathbf{\Omega}_i$, we can finally obtain a regional feature $\mathbf{S}_i$ encoding rich regional contextual information, written as.

$$\mathbf{S}_i = \delta(\mathbf{X}, \eta) \times \mathbf{\Omega}_i \tag{5}$$

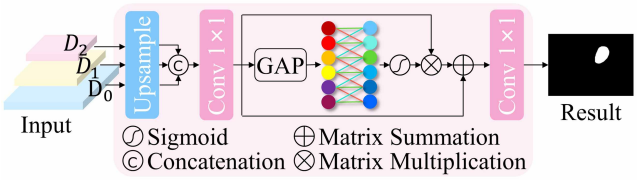

Figure 3: Our APF module. Based on the global average pooling followed by an efficient channel attention mechanism, we achieve a more effective feature fusion by adaptively reweighting multi-level output feature maps, which enables us to obtain a more refined dense prediction.

where $\mathbf{S}_i \in \mathbb{R}^{\frac{C}{4}\times(H\times W)}$. $\delta(\cdot)$ denotes a 1×1 convolution and η is the relevant learnable parameter.

Pyramid Aggregation Given the representative regional features, we can fuse them to capture different contextual information. Different from establishing long-range dependencies based on pixel-aware correlations, our regional-based method also obtains a better anti-noise ability. For an efficient pyramid aggregation, we also reshape each $\mathbf{S}_i$ to $\mathbb{R}^{C\times H\times W}$ as the same dimension with the original feature map X. By concatenating all regional features and the original feature map $\mathbf{X}$, we can obtain the final output features $\mathbf{E}$ as follow,

$$\mathbf{E} = CONCAT(\mathbf{X}, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3) \tag{6}$$

where $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$ are the three extracted regional features. After pyramid aggregation of multiple branch regional features, we enhance the feature discrimination with regional-based long-range dependencies, which can improve the accuracy of ultrasound nerve segmentation.

Adaptive Pyramid Fusion Module

Based on the high-level feature maps extracted in our RPA enhanced encoder, we can further leverage a decoder network to restore a dense prediction with the same spatial resolution as the input image. Although skip connections are widely used to obtain better precise locations, some spatial detailed information may still be lost, which eventually makes a misleading for the following denser prediction in the decoder. Because ultrasound images usually have a large amount of speckle noises, above misleading problem of detailed location information lost in the skip connections can severely affect the accuracy of our ultrasound nerve segmentation. To address above problem, multi-level feature fusion strategies (Lin et al. 2017; Xie et al. 2020) were applied and verified to be effective in improving segmentation performance. However, due to the semantic gaps existing among output feature layers, traditional indiscriminate feature fusion (Zhang et al. 2018b) based on a simple concatenation and element-wise summation still cannot obtain a satisfied segmentation accuracy, especially for a much more challenging ultrasound nerve segmentation.

Inspired by (Wang et al. 2019), to achieve a more effective output feature fusion in our decoder, we further design an adaptive pyramid fusion (APF) module to address this issue. Given pyramid feature maps $\mathbf{D}_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ from the decoder, we upsample and align them to the same dimension of $\mathbb{R}^{C\times H\times W}$, where $i \in \{0, 1, 2\}$, H, W and C are the height, width and channel number. By concatenating the three feature maps, we can obtain a larger feature map $\mathbf{F}$, which is also fed into a 1×1 convolutional function ψ for channel dimension reduction as following,

$$\mathbf{F} = \psi(CONCAT(\mathbf{UP}(\mathbf{D}_1), \mathbf{UP}(\mathbf{D}_2), \mathbf{UP}(\mathbf{D}_3)) \tag{7}$$

where $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$. $\mathbf{Up}(\cdot)$ denotes the bilinear upsampling. Moreover, we further employ an efficient channel attention mechanism to explicitly model the contextual relationship among channels, which can be expressed as following,

$$\mathbf{Y} = \mathbf{F} \oplus (\mathbf{F} \otimes \delta(f(g(\mathbf{F}), w))) \tag{8}$$

where $g(\mathbf{X}) = \frac{1}{N}\sum_{p=1}^{H}\sum_{q=1}^{W}\mathbf{X}_{(p,q)}$ is the channel-wise global average pooling for context modeling. $N = H \times W$ is the total number of pixels. $f(\cdot)$ denotes the transformation to capture channel-wise dependencies, where w is the relevant parameters. Here, the transformation is implemented by two 1D convolutions. $\delta(\cdot)$ is a Sigmoid function to generate channel attention weighting maps. After applying an element-wise multiplication to reweight the $\mathbf{F}$, we can finally obtain a more refined dense prediction $\mathbf{Y}$ based on a broadcast element-wise addition.

Loss Function

By checking each pixel one by one and comparing the predicted result for each pixel category with a label vector, cross entropy loss becomes the most commonly used loss function for image semantic segmentation tasks. As our segmentation result has only two classes, we can employ a binary entropy loss as a major component in our loss function,

$$\mathcal{L}_{bce} = \frac{1}{N} \times \sum_{i=1}^{N}(-y_i \cdot log\,(x_i) - (1 - y_i)\,log\,(1 - x_i)) \tag{9}$$

where N is the total pixel numbers. $x_i \in [0, 1]$ and $y_i \in \{0, 1\}$ denotes the predicted probability and ground truth label respectively. Considering that the distributions of nerve and other tissues in our segmentation result can be very irregular, to minimize the bias, we also use a Dice loss (Milletari, Navab, and Ahmadi 2016) as a complementary component in our loss function,

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{10}$$

where N is the total pixel numbers. $p_i \in [0, 1]$ and $g_i \in \{0, 1\}$ denotes the predicted probability and the corresponding ground truth respectively.

Finally, we obtain an overall loss function $\mathcal{L}_{total}$ with three elements, including a binary cross entropy loss $\mathcal{L}_{bce}$, a Dice loss $\mathcal{L}_{dice}$, and a L_2 regularization term as following,

$$\mathcal{L}_{total} = \mathcal{L}_{bce} + \mathcal{L}_{dice} + \frac{\lambda}{2}\|\omega\|_2^2 \tag{11}$$

where ω represents the network parameters, and λ is the decay rate. In our experiment, we set $\lambda = 0.001$.

Experiments

Dataset and Evaluation Metric

To evaluate the effectiveness of our proposed method, we conducted the experiments on the Kaggle ultrasound nerve segmentation challenge [1]. This dataset is consists of 11143 ultrasound images with a resolution of 580×420, which are manually annotated by clinical experts to generate mask images. Among the 11143 samples, 4508 and 1127 images are used for training and validation respectively, while the rest images are used for testing. To avoid overfitting and improve the generalization of our model, we also perform 4 kinds of data augmentations to enlarge the numbers of the images, including horizontal flipping, vertical flipping, diagonal flipping, and a random rotation with a degree of [-15, 15]. All images are resized to a uniform resolution of 512×512 as the input for model training. In order to sufficiently demonstrate the effectiveness of our method, we utilized the official defined training set to train our models and evaluate on the validation set. Meanwhile, the leaderboard score of our method is also collected in the website of Kaggle ultrasound nerve segmentation challenge.

This competition is evaluated on the mean Dice coefficient (DC), which is calculated by comparing the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. The leaderboard score in challenge websiet is the mean of the DC for each image in the test set. In addition, to comprehensively compare the performance of ultrasound nerve segmentation, we also employed other four additional evaluation metrics in our experiments, including Accuracy (AC), Sensitivity (SE), Specificity (SP) and Area Under the Curve (AUC).

Implementation Detail

We implemented our network by Pytorch (Paszke et al. 2017) on a 1 NVIDIA GeForce RTX 2080TI (11GB memory). In our experiments, we used ResNet-34 as our backbone, which is pre-trained on ImageNet (Russakovsky et al. 2015). During the training process, our initial learning rate is 0.0001. To obtain a more smooth convergence curve, our learning rate is also multiplied by $\left(1 - \frac{iter}{total}\right)^{power}$ with power = 0.9 after each iteration (Krogh and Hertz 1991). To speedup the network convergence, we also employed the Adam algorithm to optimize the training process. Considering that almost half of the images are background images, we also applied each mini-batch (batch size = 8) with a ratio of 1:1 of negative samples to positive samples to train our model. Our model can be converged after 70 epoches in our experiments. In addition, we also adopt the test time augmentation (TTA) (Dai et al. 2016; Zhang et al. 2018a) to improve model performance and reduce generalization error when testing all model.

Ablation Studies

To justify the advantages of our proposed method, we conducted the following ablation studies on Kaggle ultrasound nerve validation set to evaluate the effectiveness of RPA and APF modules. In our ablation studies, we used the U-shape structure equipped with a Resnet-34 backbone as the Baseline method. Two competitors were implemented (Baseline+RPA and Baseline+APF) by adding RPA or APF module to the Baseline respectively. Finally, the RPA and APF modules are simultaneously added to the Baseline (Baseline+RPA+APF).

[1] https://www.kaggle.com/c/ultrasound-nerve-segmentation

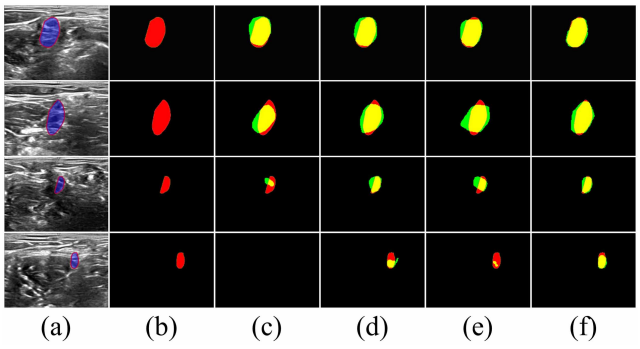

Figure 4: Visual comparison of our ablation studies. (a) Input image. (b) Ground truth. (c) Baseline. (d) Baseline+RPA. (e) Baseline+APF. (f) Ours (Baseline + RPA + APF). Green and red pixels indicate the predictions and ground truth respectively. Yellow pixels represent the overlap regions between the prediction and ground truth.

Method	DC (%)	ACC (%)	AUC (%)
Baseline	70.71	99.01	98.77
Baseline+RPA	73.34	99.29	99.24
Baseline+APF	72.32	99.18	99.11
Baseline+RPA+APF	**74.23**	**99.37**	**99.66**

Table 1: Statistical comparison of our ablation studies.

Ablation Study for RPA Module As the highest level semantics is mainly contained in the deepest layer, our RPA module embedded on the top layer can exploit the most representative regional context information. Visual comparison of our ablation studies is as shown in Figure 4, where several typical challenging cases with various scales and extremely low contrast boundaries can be seen. Without the enhancement of RPA, we can clearly observe that Baseline would be easily failed in segmenting the small nerve structures, especially when the boundaries of small nerves are also blur. By enhancing the feature discrimination with regional-based long-range dependencies, Baseline+RPA significantly improves the segmentation accuracy, even where have large scale variations and irregular color contrast near the boundaries of nerves, which clearly demonstrate the effectiveness of our RPA module in addressing challenging ultrasound nerve segmentation with regional multi-scale contexts.

In addition, we also performed a statistical comparison in our ablation studies by collecting the mean DC, ACC and AUC values over the Kaggle ultrasound nerve validation set. As shown in Table 1, the Baseline+RPA method achieves

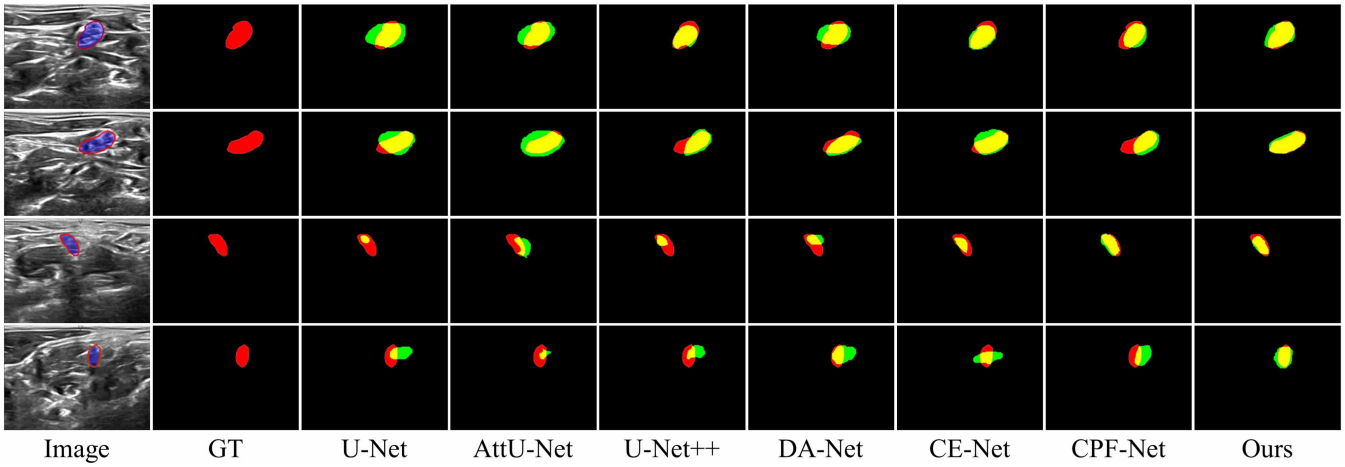

Figure 5: Visual comparison with different state-of-the-art methods in ultrasound nerve segmentation. Green and red pixels indicate the predictions and ground truth respectively. Yellow pixels represent the overlap regions between the prediction and ground truth.

Method	Year	DC (%)	ACC (%)	SE (%)	SP (%)	AUC (%)
U-Net	2015	70.27	98.98	84.44	99.38	98.73
AttU-Net	2018	71.42	99.09	84.91	99.43	98.86
U-Net++	2018	72.03	99.15	86.42	99.44	99.03
DA-Net	2019	73.28	99.26	**86.68**	99.48	99.26
CE-Net	2019	72.74	99.21	85.62	**99.51**	99.18
CPF-Net	2020	73.48	99.29	86.25	99.49	99.31
Ours	2020	**74.23**	**99.37**	86.55	**99.51**	**99.66**

Table 2: Statistical comparison with state-of-the-art methods on the validation set.

73.34%, 99.29% and 99.24% in DC, ACC and AUC metrics, outperforming the Baseline by 2.63%, 0.28% and 0.5% respectively. This also obviously points out that our RPA module effectively improve the ultrasound nerve segmentation by fully exploiting of more discriminative contexts based on regional long-range dependencies.

Ablation Study for APF Module As mentioned before, ultrasound images usually have a large amount of speckle noises, which produces a difficult problem for the traditional multi-level feature fusion only utilizing simple concatenation. Instead, by adaptively reweighting multi-level output feature maps based on an efficient channel attention mechanism, our APF module can achieve a more refined dense prediction. Our ablation study also validated the effectiveness of our APF in both visual and statistical comparisons. As shown in Figure 4 (e), we can clearly observe that the yellow overlap regions in Baseline+APF is much larger than the Baseline, indicating that our APF obtained a better nerve segmentation relying on the adaptive multi-level feature fusion. On the other hand, the superiority of our APF also obviously demonstrated in the statistical results shown in Table 1. Compared with the Baseline, Baseline+APF achieves

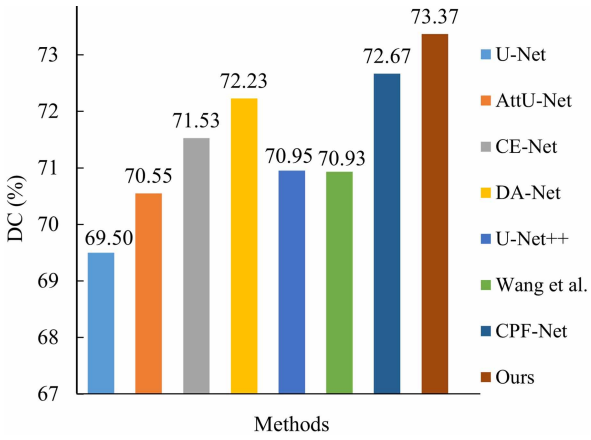

Figure 6: Statistical comparison on DC metric with different state-of-the-art methods on the challenge website.

1.61%, 0.17% and 0.34% accuracy improvement in terms of DC, ACC and AUC metrics, respectively.

Finally, based on a seamless combination of both RPA and APF modules, Baseline+RPA+APF not only effectively distinguishes the target nerve regions from the blur background areas based on a novel region-level attention mechanism, but also refine the final dense prediction by meticulously and adaptively fusing multi-level feature maps, as shown in the Figure 4 (f) and Table 1.

Comparison with State-of-the-art Methods

To further validate the superiority of our model, we also conducted a comparison among our method and several state-of-the-art methods, including U-Net (Ronneberger, Fischer, and Brox 2015), Attention U-Net (Oktay et al. 2018), U-Net++ (Zhou et al. 2018), DA-Net (Fu et al. 2019), CE-Net (Gu et al. 2019), and CPF-Net (Feng et al. 2020). In our experiments, we implemented all competitors on the same

computing environments and with the same data augmentations to guarantee a fair comparison.

Typical challenging ultrasound nerve images and the segmentation results for different competitors are as shown in Figure 5, where we can clearly observe the low gray contrast, irregular shapes, scale variations and non-uniform distributions of nerves in the ultrasound images. Only relying on simply stacking continuous convolution and pooling operations, U-Net still cannot obtain a stable segmentation performance, especially for the cases with scale variation or low gray contrast. By combining U-Nets to assemble deeper network, U-Net++ obtained some accuracy improvements but is still unstable when processing challenging cases. CE-Net also further improve the segmentation accuracy by capturing more high-level information and retaining richer spatial information. On the other hand, to tackle blurred boundaries caused by low gray contrast, which is commonly seen in ultrasound images, Attention U-Net also tried to utilize an attention gate to suppress irrelevant background noises, which is still difficult to accurately segment the small ultrasound targets. Similarly, based on both spatial attention and channel attention, DA-Net further obtained a better nerve segmentation performance. More recently, by exploiting rich long-range contexts followed by a progressive feature fusion, CPF-Net achieved better segmentation results than the above methods. However, all the above competitors neither fully utilized regional level representation model to capture rich regional context information, nor meticulously and adaptively fused the multi-level feature maps to refine a dense prediction. Therefore, none of them can obtain accurate and robust enough performance to satisfy the challenging ultrasound nerve segmentation task, commonly appearing with low gray contrast, irregular shapes, and various scales and target distributions. As shown in Figure 5, our model obtains the best segmentation results compared with other methods, where segmented nerves are the closest to the ground truth. The visual experimental results clearly demonstrated that our proposed methods can not only accurately segment the nerves with various scales and irregular shapes, but also better refine the boundaries where have low gray contrast between nerve and surrounding tissues.

In addition, we also performed a statistical comparison with state-of-the-art methods by collecting the mean DC, ACC, SE, SP and AUC values over the Kaggle ultrasound nerve dataset. From the results shown in Table 2, we can clearly see that our model also generally outperforms other competitors by achieving 74.23% 99.37%, 86.55%, 99.51%, 99.66% in DC, ACC, SE, SP and AUC metrics respectively. On the other hand, we also compared our method with other competitors on the test set by submitting results of different methods to the official challenge website. We plotted the histogram graph to visualize the accuracy comparison among our network and the other 6 competitors. As shown in the Figure 6, we can also clearly see the superiority of our method in the ultrasound nerve segmentation for the test cases in Kaggle dataset. Even compared with (Wang, Shen, and Zhou 2019), which is the latest published work about Kaggle ultrasound nerve segmentation, our model still achieves about 3% improvement in the major DC metric,

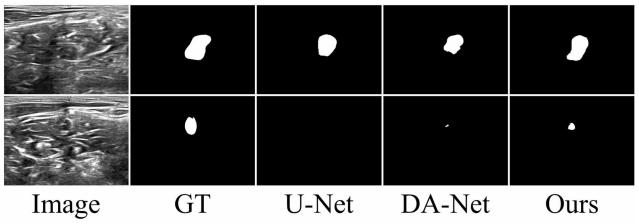

Figure 7: Failure cases for the competitors and our model.

clearly pointing out the effectiveness of our RPA and APF modules in addressing challenging issues in ultrasound nerve segmentation.

Discussions and Limitations

Through the above ablation studies and comparative experiments, we have observed that even if there are many complex nerve structures with various scales and low contrast, our method still successfully achieves satisfactory results owing to the proposed RPA and APF modules. By utilizing the self-attention mechanism to encode the region-wise features based on the pyramid pool sampling strategy, our RPA module can not only captures multi-scale contextual semantic information, but also obtain more discriminative regional features. Moreover, the idea of our proposed RPA also provides a good hint for segmenting objects from ultrasound images, where pixel-wise modeling may inevitably introduce much speckle noises in the computation and potentially degrade segmentation performance. By dynamically selecting and fusing the important and dominant output features based on an efficiency channel attention mechanism, our APF further successfully refines the final dense prediction, which also provides a good tool for optimizing the segmentation networks where multi-level output feature fusion is required to address large scale variation among shallow layers in the decoder. On the other hand, our method still has some limitations. As shown in Figure 7, our method still fails to segment the cases with extremely small scales, as well as extremely low contrast near the boundaries between the target and surrounding tissues.

Conclusion

We presented a novel deep network for automatic nerve segmentation from ultrasound images, which is a quite challenging considering the large variation of targets and the low image quality. A novel region-aware global context modeling techniques are proposed to establish long-range dependencies in a more efficient way than traditional pixel-wise modeling techniques. An adaptive pyramid fusion scheme is developed to meet the spatial-semantic gap between high-level and low-level features and exclude background noise when performing fusion. Extensive experiments demonstrate the superiority of the proposed network to our rivals. Future investigations include testing it on more ultrasound datasets and integrate it in ultrasound-guided interventions.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of China under Grant 61973221, the Natural Science Foundation of Guangdong Province of China under Grant 2018A030313381 and Grant 2019A1515011165, the Key Lab of Shenzhen Research Foundation of China under Grant 201707311550233, the COVID-19 Prevention Project of Guangdong Province of China under Grant 2020KZDZX1174, the Major Project of the New Generation of Artificial Intelligence of China under Grant 2018AAA0102900, and the Hong Kong Research Grants Council under Grant 15205919.

References

Baby, M.; and Jereesh, A. 2017. Automatic nerve segmentation of ultrasound images. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 1, 107–112.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, 1971–1980.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR* abs/1606.00915.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4): 834–848.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 379–387.

Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; and Chen, X. 2020. CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation. *IEEE Transactions on Medical Imaging* 1–1.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual Attention Network for Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 3146–3154.

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Trans. Medical Imaging* 38(10): 2281–2292.

Hadjerci, O.; Hafiane, A.; Conte, D.; Makris, P.; Vieyres, P.; and Delbos, A. 2016. Computer-aided detection system for nerve identification using ultrasound images: A comparative study. *Informatics in Medicine Unlocked* 3: 29–43.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2020. IAUnet: Global Context-Aware Feature Learning for Person Re-Identification. *CoRR* abs/2009.01035.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. CCNet: Criss-Cross Attention for Semantic Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 603–612.

Kakade, A.; and Dumbali, J. 2018. Identification of nerve in ultrasound images using u-net architecture. In *2018 International Conference on Communication information and Computing Technology (ICCICT)*, 1–6.

Krogh, A.; and Hertz, J. A. 1991. A Simple Weight Decay Can Improve Generalization. In Moody, J. E.; Hanson, S. J.; and Lippmann, R., eds., *Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991]*, 950–957.

Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; and Liu, H. 2019. Expectation-Maximization Attention Networks for Semantic Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9166–9175.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 936–944.

Liu, C.; Liu, F.; Wang, L.; Ma, L.; and Lu, Z.-M. 2018. Segmentation of nerve on ultrasound images using deep adversarial network. *Int. J. Innov. Comput. Inform. Control* 14(1): 53–64.

Liu, W.; Rabinovich, A.; and Berg, A. C. 2015. ParseNet: Looking Wider to See Better. *CoRR* abs/1506.04579.

Luo, W.; Li, Y.; Urtasun, R.; and Zemel, R. S. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 4898–4906.

Milletari, F.; Navab, N.; and Ahmadi, S. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, 565–571.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M. C. H.; Heinrich, M. P.; Misawa, K.; Mori, K.; McDonagh, S. G.; Hammerla, N. Y.; Kainz, B.; Glocker, B.; and Rueckert, D. 2018.

Attention U-Net: Learning Where to Look for the Pancreas. *CoRR* abs/1804.03999.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N.; Hornegger, J.; III, W. M. W.; and Frangi, A. F., eds., *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, 234–241.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115(3): 211–252.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 1–9.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 5998–6008.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2019. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *CoRR* abs/1910.03151.

Wang, R.; Shen, H.; and Zhou, M. 2019. Ultrasound Nerve Segmentation of Brachial Plexus Based on Optimized ResU-Net. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–6.

Wang, X.; Girshick, R. B.; Gupta, A.; and He, K. 2018. Non-Local Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7794–7803.

Xie, Q.; Lai, Y.; Wu, J.; Wang, Z.; Zhang, Y.; Xu, K.; and Wang, J. 2020. MLCVNet: Multi-Level Context VoteNet for 3D Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10444–10453.

Zhang, Y.; Qiu, Z.; Yao, T.; Liu, D.; and Mei, T. 2018a. Fully Convolutional Adaptation Networks for Semantic Segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6810–6818. IEEE Computer Society.

Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; and Sun, J. 2018b. ExFuse: Enhancing Feature Fusion for Semantic Segmentation. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, 273–288.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid Scene Parsing Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6230–6239.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Stoyanov, D.; Taylor, Z.; Carneiro, G.; Syeda-Mahmood, T. F.; Martel, A. L.; Maier-Hein, L.; Tavares, J. M. R. S.; Bradley, A. P.; Papa, J. P.; Belagiannis, V.; Nascimento, J. C.; Lu, Z.; Conjeti, S.; Moradi, M.; Greenspan, H.; and Madabhushi, A., eds., *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*, volume 11045 of *Lecture Notes in Computer Science*, 3–11.