Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 4, No. 6, 2024

Medical Image Segmentation with Bilateral Spatial Attention and Transfer Learning

Dan Sun[1], Mingxiu Sui[2], Yingbin Liang[3] , Jiacheng Hu[4] , Junliang Du[5]

[1]Washington University in St. Louis, St. Louis, USA

[2]University of Iowa, Iowa City, USA

[3]Northeastern University, Seattle, USA

[4]Tulane University, New Orleans, USA

[5]Shanghai Jiao Tong University, Shanghai, China

Correspondence should be addressed to Junliang Du; jldu@acm.org

Abstract:

In the medical field, image processing technology is becoming more and more extensive, especially in disease screening and diagnosis. However, traditional medical image processing methods can produce a lot of noise and artifacts in low-dose X-ray CT images, affecting a doctor's diagnostic accuracy. At the same time, early lesions, such as colon polyps, is difficult, and existing image segmentation algorithms struggle to locate and separate lesion areas accurately. With the development of deep learning technology, especially the application of attention mechanisms in medical image processing, new solutions are provided. Attention mechanisms have shown excellent performance in noise suppression and the accurate segmentation of lesion tissue, but they rely on a large amount of training data. The lack of medical image data limits their application. In order to solve these problems, this paper proposes a medical image denoising and segmentation algorithm that combines transfer learning and attention mechanisms. We design an integrated medical image auxiliary diagnosis system based on this algorithm to improve the efficiency and accuracy of medical image processing.

Keywords:

Medical image processing; image segmentation; deep learning; attention mechanism

1. Introduction

In the medical field, image segmentation methods provide doctors with clear and intuitive visual representations, helping them identify patients' internal structures and disease states. They are becoming increasingly important in disease screening. However, the application of image processing algorithms in the clinical field has gradually become more widespread, and at the same time, it has exposed many thorny technical problems. Although the low-dose computed tomography (CT) used in hospitals and clinics reduces the potential harm of radiation to patients, it also produces a large amount of noise. Current algorithms cannot completely remove this interference, which blurs the original tissue structure and affects subsequent clinical judgments. At the same time, in certain medical images, such as colon endoscopy, early polyp lesion areas are subtle and indistinct. They are usually connected to the surrounding mucosal tissue, and existing methods cannot locate and distinguish them, making it difficult to maintain a high level of machine recognition accuracy. In recent years, with the development of deep learning, especially the rise of visual attention mechanisms, new solutions have been brought about. These mechanisms are demonstrated when dealing with long-distance dependencies under various medical images and solving the balance problem between local and global information. They have unique advantages and have outstanding performance in suppressing noise and accurately segmenting diseased tissues. However, the attention mechanism relies heavily on the training of large-scale data images,

making the problem of the lack of medical image data sets increasingly prominent, hindering further development of attention mechanisms in this area.

Based on the above issues and considerations, this article introduces the idea of transfer learning to apply and improve the attention mechanism, designing a medical image segmentation algorithm. On this basis, we create an original design of a medical image auxiliary diagnosis system to achieve integrated processing of medical image segmentation. The specific research content of this article is as follows:

(1) To address the problem of inaccurate segmentation and positioning in current methods due to the diverse morphology and blurred boundaries of tissues or lesions in medical images such as colonoscopy, this paper proposes a bilateral spatial attention segmentation network from the perspective of designing multiple attention-related feature learning. The bilateral spatial attention module is used to locate difficult-to-distinguish tissues or lesion, effectively capturing the correlation between lesion or tissue features to achieve mutual complementation of useful information.

(2) The attention rendering module optimizes the edge contour details to accurately predict the output and obtain an accurate segmented image. The algorithm has been extensively experimented on two types of public datasets and verified using a variety of evaluation indicators. The results show that it outperforms advanced comparative methods in both qualitative and quantitative performance.

2. Related work

Recent progress in deep learning, particularly convolutional neural networks (CNNs) and attention mechanisms, has made significant strides in improving medical image analysis, including segmentation and denoising tasks. These advancements provide critical foundations for the development of accurate and efficient image segmentation frameworks, especially in scenarios like low-dose CT scans and colonoscopy imaging, where noise and subtle lesion boundaries pose significant challenges.

Several studies have contributed to this domain. Xiao et al. [1] employed a CNN-based approach for classifying cytopathology images, showcasing the network's ability to extract essential features in a medical context. Their work underlines the strength of CNNs in handling complex medical image data, aligning with our use of convolutional layers for initial feature extraction before applying attention mechanisms. This foundational work on CNNs helps to inform the early layers of our segmentation network, ensuring robust representation of medical features.

The challenge of registering and segmenting medical images has also been addressed in the context of deep learning models augmented by attention mechanisms. Liang et al. [2] introduced an unsupervised image registration method using Dense U-Net combined with channel attention, which improves image alignment and segmentation accuracy. Their work directly relates to our method as we similarly incorporate attention mechanisms to capture subtle spatial features. By building on these concepts, our bilateral spatial attention mechanism effectively enhances feature learning, especially in noisy or complex medical images where lesions are difficult to distinguish.

Optimization techniques also play a critical role in improving the performance of deep learning models. Qin et al. [3] proposed the RSGDM (Reduced-Stochastic Gradient Descent Momentum) approach to address biases in optimization, providing a novel method that could benefit medical image segmentation. Optimization is crucial for ensuring accurate convergence in deep learning models, and integrating effective optimization methods like RSGDM could further refine the segmentation results in our network, particularly when training on smaller medical datasets, as addressed through our use of transfer learning.

In the realm of feature representation, Yan et al. [4] presented methods for spatiotemporal feature representation in time-series data. Their work is significant in demonstrating how complex, multidimensional data can be effectively processed and segmented. This informs our approach in handling spatial information in medical images, where capturing long-range dependencies through attention mechanisms is vital for accurate segmentation of irregular lesions.

Yao et al. [5] highlighted the advantages of transformer models in maintaining depth gradient continuity in image processing tasks. Their findings support the argument that attention-based models, like the one used in our bilateral spatial attention module, can improve segmentation performance by maintaining a

balance between local and global feature extraction. This continuity is essential in medical imaging to ensure that lesion boundaries are accurately captured.

While Wu et al. [6] and Zhang et al. [7] focus on graph neural networks and contrastive learning in other domains, their exploration of syntactic features and contrastive methods presents potential avenues for further refinement of feature learning in medical imaging. These approaches can inspire future enhancements to attention mechanisms and transfer learning in medical image segmentation by improving the generalization capabilities of deep learning models.

By building on these foundational works, our proposed bilateral spatial attention segmentation network, combined with transfer learning, addresses the limitations of traditional segmentation methods in handling noisy medical images and imprecise lesion boundaries. Specifically, our method enhances feature extraction and segmentation accuracy by leveraging bilateral spatial attention to focus on both local and global features, optimizing segmentation even with limited training data.

3. Algorithm principle

Figure 1 shows the framework of the bilateral spatial attention medical image segmentation network designed in this chapter. The network is an encoder-decoder U-shaped structure. By integrating multi-scale features, the segmentation network can improve its ability to recognize and segment objects of different scales, thereby improving the accuracy, robustness, and generalization of the segmentation results. After the medical image is input, the ResNeXt [8] module is first used to replace the convolution block as part of the encoder to perform preliminary feature extraction. ResNeXt is an extension and improvement of the residual network. It adopts a grouped convolution structure. By adding groups in the convolution layer, the network can obtain more representation capabilities without increasing parameters and computational complexity, thereby improving the performance of the network. After preliminary feature extraction, the feature map of the image is obtained. For the convenience of description, the subsequent feature map will be referred to as f_i.

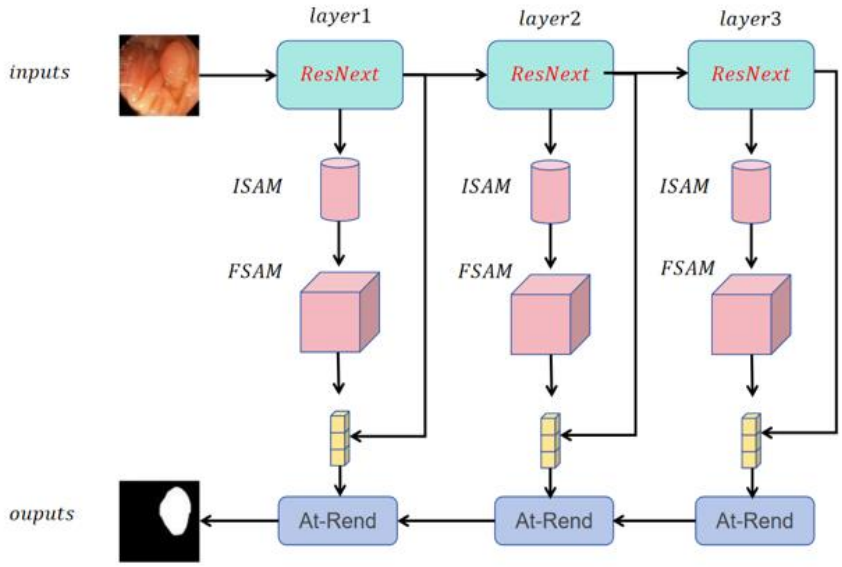

Figure 1. Framework of the bilateral spatial attention medical image segmentation network

For subtle details such as fine tissues and lesions in the image that are not easily perceived, the network needs to pay more attention to the high-level semantic information of the local area, to improve the network's perception of local details. To effectively capture local features and global context information in each feature map to locate important areas, this paper designs an image space attention module (ISAM). As shown in Figure 2, it consists of 6 Swin Transformer modules (STB), a convolution module (Conv), and a residual connection, where each convolution module is a combination of a two-dimensional 3×3 convolution layer, batch normalization (Batch Norm), and ReLU activation function. Drawing on the experience of designing the model in Chapter 3, since large-scale deep learning methods such as Transformer still require large-scale datasets for sufficient training, this chapter will continue to use the

idea of transfer learning to use the Swin Transformer pre-trained model in natural images for the training of this model. Given the input feature $f_{i,0}$ of the image space attention module, the intermediate features are first extracted through multiple stacked Swin Transformer modules. The formula is defined as follows:

$$f_{i,j} = STB(f_{i,j-1})$$
$$i = 1,2,...$$
$$j = 1,2,...$$

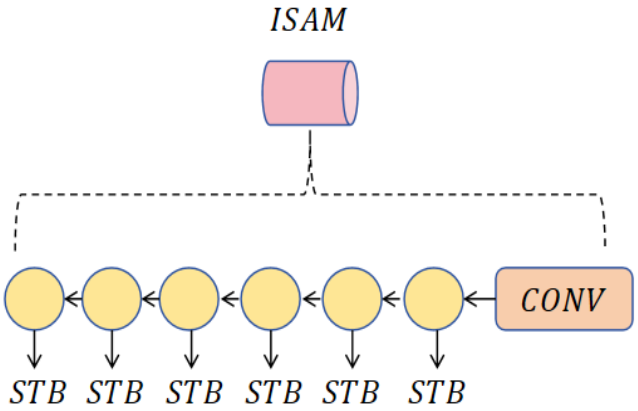

Figure 2. Image-space attention module

$STB_{i,j}(\cdot)$ Represents the processing of the image by the Swin Transformer block in the image space attention module. Then, the convolution block is connected and added to the input features of the module to obtain the output formula of the module:

$$f_i = Conv(f_i) + f_{i,0}$$

Where $Conv(\cdot)$ refers to the operation process of the convolution block in the image spatial attention module.

When designing the convolutional blocks in the image space attention module, it is necessary to consider how to refine the input feature map into a richer feature representation so that the subsequent feature space attention module can more easily capture important information. Here, a convolutional block composed of operations such as convolution will be used to achieve this goal. Residual connections can be achieved by weighted addition of the input features to the output of the convolutional block, thereby effectively aggregating and fusing features at different levels. Adding residual connections at the beginning and end of the module enables the network to better retain and transmit the original input information.

To capture meaningful objective features from different receptive fields and further improve the robustness of model representation, this paper designs a feature space attention module (FSAM) to guide the network model to adaptively learn more feature representations. This module complements the image space attention module.

In this module, the feature map is first input into the convolution block (Conv) and the dilated convolution block (DConv). Generally, the convolution operation process has a small receptive field, while the dilated convolution expands the receptive field by increasing the dilation rate of the convolution kernel, thereby being able to capture a wider range of spatial information[9]. By connecting the convolution block and the dilated convolution block in series, a feature extraction network with different receptive fields can be constructed to efficiently capture long-range dependencies. The dilated convolution block contains a dilated convolution layer (with a dilation rate of 2), a batch normalization layer, and a ReLU activation function.

$$f_{gap} = GAP(DConv(Conv(f_i)))$$

$$f_{gmp} = GMP(DConv(Conv(f_i)))$$

Among them, $DConv(\cdot)$ represents the dilated convolution block operation, $GAP(\cdot)$ and $MAP(\cdot)$ represent the global average pooling operation and the global maximum pooling operation. The two operations are complementary for extracting different types of feature information. Global average pooling can extract the overall feature information, while global max pooling can capture the most significant local features. The combination of the two can increase the feature space information of the feature map. After obtaining the feature space information, a more representative feature map is adaptively obtained from the receptive field. The formula is as follows:

$$f_{gap}^{'} = H_{FC}(f_{gmp})$$
$$f_{gmp}^{'} = H_{FC}(f_{gmp})$$

Where represents a series of operations of the fully connected layer. Finally, the two feature vectors are merged and element-wise summed, and then the sigmoid activation function is used to obtain the feature space attention map α. Multiplying α with the input feature map can achieve automatic selection of enhanced features. The following two formulas describe the process of performing feature space attention:

$$\alpha = \sigma(f_{gap}^{'} + f_{gmp}^{'})$$
$$f_i^{\alpha} = \alpha \times f_i$$

Where $\sigma(\cdot)$ represents the sigmoid activation function. After the residual connection, the final output of the entire module is expressed as:

$$f_i = f_i^{\alpha} + Conv(f_i)$$

There are two main purposes of connecting the feature space attention module in series after the image space attention module: first, it enables the network to pay more attention to the image position information to enhance the capture of key details while minimizing the interference of image space information on the correlation between feature maps; second, the feature space attention module can reduce the dimension of the feature map, which can greatly reduce the computational complexity.

4. Experimental design

4.1. Experimental setup

For the polyp segmentation experiment, the ratio of the training set, validation set, and test set is divided into 7:1:2. The training set and validation set are randomly selected from Kvasir, CVC-ColonDB, and CVC-EndoScene, and the test set is selected from the remaining Kvasir, CVC-EndoScene, and ETIS images. The model is trained on the training sets of ISIC 2016[10] and ISIC 2018[11], and evaluated on PH2[12] and the remaining datasets images. All experiments do not perform any data augmentation before training. The deep learning framework uses PyTorch to complete the training process, and the training process will be run on two NVIDIA RTX 3090 GPUs with 24G video memory. Since the image resolutions of the datasets vary, they are resized to 256×256 resolution for training and testing before training. The batch size is set to 16, and the Adam optimizer is used with a learning rate of 0.001. The pre-trained weights on the ImageNet[13] dataset are used to initialize the network parameters of the Swin Transformer block. The total number of network parameters is 6.36M. The training takes a total of 11 hours, and the model is fitted after about 120 epochs.

4.2. Datasets

This section will conduct experimental verification and analysis on four public and authoritative colonoscopy image segmentation datasets and two skin image datasets. For polyp segmentation, experiments were conducted on the Kvasir[14], ETIS[15], CVC-ColonDB[16], and CVC-EndoScene[17] datasets. The Kvasir dataset contains 1,000 images taken and annotated by endoscopists at Oslo University Hospital in Norway. The ETIS dataset consists of 192 images and is the official dataset released by the International Medical Image Computing and Computer-Assisted Intervention Association

(MICCAI). ColonDB is generated from 300 images randomly cut from colorectal examination videos of the Mayo Clinic of the National University of the United States. CVC-EndoScene contains two datasets, which include 612 images of ClinicDB and 60 images of CVC-300 collected from 29 colonoscopy video sequences.

4.3. Experimental Results

To demonstrate the advancement of our method, this section selects several mainstream polyp segmentation methods. The following methods are selected based on code availability and relevance to the proposed method, including U-Net, Unet++, PraNet, SANet, and LDNet.

Table 1: Kavsir Dataset Experimental Results

Kavsir						
Model	mDSC	mIOU	S_α	F_β^ω	$E_\varphi^{\max}$	MAE
U-Net	0.832	0.739	0.754	0.746	0.822	0.072
Unet++	0.841	0.718	0.760	0.870	0.893	0.046
PraNet	0.880	0.865	0.845	0.926	0.960	0.034
SANet	0.882	0.823	0.871	0.889	0.954	0.042
LDNet	0.890	0.877	0.866	0.876	0.936	0.070
Ours	0.901	0.887	0.885	0.917	0.973	0.029

Table 2: ETIS Dataset Experimental Results

ETIS						
Model	**mDSC**	**mIOU**	S_α	F_β^ω	$E_\varphi^{\max}$	**MAE**
U-Net	0.608	0.462	0.471	0.704	0.735	0.065
Unet++	0.691	0.579	0.587	0.789	0.786	0.052
PraNet	0.775	0.713	0.710	0.834	0.863	0.030
SANet	0.798	0.721	0.701	0.828	0.835	0.044
LDNet	0.746	0.722	0.726	0.823	0.865	0.037
Ours	0.833	0.749	0.729	0.842	0.894	0.024

Table 3: CVC-ColonDB Dataset Experimental Results

CVC-ColonDB						
Model	**mDSC**	**mIOU**	S_α	F_β^ω	$E_\varphi^{\max}$	**MAE**
U-Net	0.404	0.340	0.513	0.754	0.770	0.066
Unet++	0.411	0.345	0.608	0.801	0.830	0.057
PraNet	0.684	0.643	0.756	0.843	0.898	0.046

SANet	0.741	0.664	0.747	0.838	0.888	0.049
LDNet	0.746	0.669	0.758	0.845	0.909	0.045
Ours	0.753	0.672	0.755	0.850	0.900	0.031

Table 4: CVC-EndoScene Dataset Experimental Results

CVC-EndoScene						
Model	**mDSC**	**mIOU**	S_α	F_β^ω	$E_\varphi^{\max}$	**MAE**
U-Net	0.800	0.773	0.776	0.898	0.895	0.035
Unet++	0.831	0.843	0.810	0.919	0.927	0.029
PraNet	0.870	0.873	0.913	0.938	0.945	0.023
SANet	0.864	0.863	0.859	0.921	0.943	0.026
LDNet	0.873	0.875	0.926	0.925	0.948	0.021
Ours	0.880	0.880	0.933	0.940	0.956	0.010

The evaluation index results of various methods on polyps are listed in Table 1 to Table 4. In the test results of the Kavsir dataset, ours performed best, ranking first in all six test indicators, followed by LDNet. It is worth noting that PraNet and SANet each surpassed LDNet in one indicator and ranked second. In the test results of the ETIS dataset, since all images in the ETIS dataset were not involved in the training, ours showed a huge advantage, which also proved that ours has strong robustness in unfamiliar test environments. In the test results of the ConlonDB dataset, the low contrast of the image caused the unclear image structure, making it easier for LDNet to play the advantages of the perception module, so the comparison performance results on the and indicators were the best, and ours was the best in the remaining four indicators. In the test results of the CVC-EndoScene dataset, ours obtained the best test results and improved in all indicators.

Based on the comparison of the above indicators, the following analysis can be drawn. As the most outstanding polyp segmentation method in MICCAI 2022, LDNet still has a certain dominance. The results on the A and MAE indicators prove that the lesion-aware cross-attention module designed by it can distinguish the feature contrast between polyps and background areas and capture long-range contextual relationships through the self-attention module to obtain accurate segmentation effects. PraNet has been the baseline method for polyp segmentation in recent years, and its performance in the indicator is relatively outstanding. It generates a global feature map by aggregating features in high-level layers in parallel partial decoders and guides the reverse attention module to mine boundary clues. Through the clever combination of small multi-level feature fusion and attention modules, SANet is also highly competitive in indicator comparison.

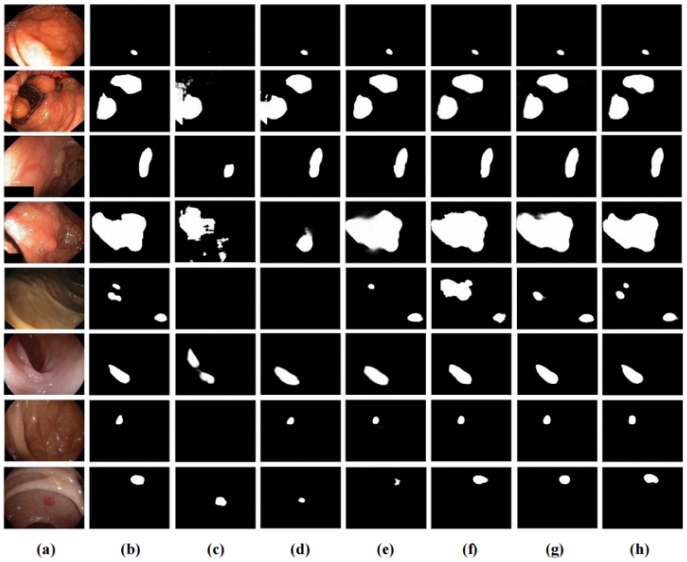

Figure 3. Visual comparison of methods on different polyp segmentation image datasets

Figure 3 shows the experimental results of our algorithm on the dataset. From the experimental results, we can see that our method has good visual characteristics and can well segment the segmentation targets in the dataset.

5. Conclusion

In order to solve the problem of inaccurate segmentation in current methods due to diverse morphologies and blurred boundaries of tissues or lesions in medical images, this paper proposes a bilateral spatial attention segmentation network, in which the Image Space Attention Module effectively captures local features and global background information by enhancing local attention weights to locate important areas in the image; the Feature Space Attention Module adaptively adjusts the feature space attention weights to give relatively critical and useful feature maps more attention. At the same time, an Attention Rendering Module is introduced, which draws on the optimization idea of rendering to ensure the generation of high-quality masks with accurate edges to improve the current segmentation performance of medical images with uncertain tissue or lesion boundaries. Finally, a series of comparative experiments and visual feature demonstrations on four public and authoritative colonoscopy image datasets are conducted to prove the effectiveness and superiority of the network designed in this paper.

References

[1] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 145-149, 2024.
[2] Y. Liang, Y. Zhang, Z. Ye, and Z. Chen, "Enhanced Unsupervised Image Registration via Dense U-Net and Channel Attention," Journal of Computer Science and Software Applications, vol. 4, no. 5, pp. 8-15, 2024.
[3] H. Qin, H. Zheng, B. Wang, Z. Wu, B. Liu, and Y. Yang, "Reducing Bias in Deep Learning Optimization: The RSGDM Approach," arXiv preprint arXiv:2409.15314, 2024.
[4] X. Yan, Y. Jiang, W. Liu, D. Yi, H. Sang, and J. Wei, "Data-Driven Spatiotemporal Feature Representation and Mining in Multidimensional Time Series," arXiv preprint arXiv:2409.14327, 2024.
[5] J. Yao, T. Wu, and X. Zhang, "Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with CNN," arXiv preprint arXiv:2308.08333, 2023.
[6] L. Wu, Y. Luo, B. Zhu, G. Liu, R. Wang, and Q. Yu, "Graph Neural Network Framework for

Sentiment Analysis Using Syntactic Feature," arXiv preprint arXiv:2409.14000, 2024.
[7] Z. Zhang, J. Chen, W. Shi, L. Yi, C. Wang, and Q. Yu, "Contrastive Learning for Knowledge-Based Question Generation in Large Language Models," arXiv preprint arXiv:2409.13994, 2024.
[8] Xie S., Girshick R., Dollár P., et al., "Aggregated residual transformations for deep neural networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5987-5995, 2017.
[9] Wang Q., Wu B., Zhu P., et al., "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11531-11539, 2020.
[10] Gutman D., Codella N. C., Celebi E., et al., "Skin Lesion Analysis toward Melanoma Detection: A Challenge hosted by the International Skin Imaging Collaboration," 2016.
[11] Codella N., Rotemberg V., Tschandl P., et al., "Skin Lesion Analysis toward Melanoma Detection: A Challenge hosted by the International Skin Imaging Collaboration," 2019.
[12] Yu F., Koltun V., "Multi-Scale Context Aggregation by Dilated Convolutions," Proceedings of the International Conference on Learning Representations, 2016.
[13] Krizhevsky A., Sutskever I., Hinton G. E., "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.
[14] Jha D., Smedsrud P. H., Riegler M. A., et al., "Kvasir-seg: A segmented polyp dataset," MultiMedia Modeling: International Conference, pp. 451-462, 2020.
[15] Silva J., Histace A., Romain O., et al., "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," International Journal of Computer Assisted Radiology and Surgery, vol. 9, no. 2, pp. 283-293, 2013.
[16] Tajbakhsh N., Gurudu S. R., Liang J., "Automated polyp detection in colonoscopy videos using shape and context information," IEEE Transactions on Medical Imaging, vol. 35, no. 2, pp. 630-644, 2016.
[17] Azquez V., Bernal D., Anchez J., et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," Journal of Healthcare Engineering, pp. 1-9, 2017.