Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 3, 2023

Mechanisms and Applications of Attention in Medical Image Segmentation: A Review

-- Subtitle Is Not Required, Please Write It Here If Your Article Has One

Yabei Li[1, *], Minjun Liang[1], Mingyang Wei[1], Ge Wang[1], Yanan Li[2]

[1] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan, 454000, P. R. China
[2] Institute of Orthopaedics and Traumatology, The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China.
[*] **Corresponding author**: Yabei Li (liyabei@home.hpu.edu.cn)

Abstract: The core task of medical image segmentation based on deep learning is to quickly obtain good results through low-cost auxiliary modules. The attention mechanism, relying on the interacting features of the neural network, is one of the lightweight schemes to focus on key features, which is inspired by the characteristics of selective filtering information in human vision. Through the investigation and analysis, this paper argues that the common attentional mechanisms can be mainly classified into four types according to their structure and form: (i) conventional attention based on feature interaction, (ii) multi-scale/multi-branch-based attention, (iii) Self-similarity attention based on key-value pair queries, (iv) hard attention, etc.
Medical images contain poor and blur descriptions of contextual information than natural images. They are usually re-imaging by the feedback intensity of the medium signal since most of them have low contrast and uneven appearance, as well as contain noise and artifacts. In models based on deep learning, without the ability to focus on key descriptive information or features, it is difficult for well-designed models to perform theoretically. This paper shows that attention mechanisms can guide downstream medical image analysis tasks to master discernible expected features while filtering and suppressing irrelevant information to enhance the intensity of target features. Therefore, the network performance can be improved through continuous highly accurate feature spatial evolution.

Keywords: Attentional mechanisms, Image segmentation, Medical image analysis.

1. Introduction

In 2012, Alex-Net [1] reduced the top-5 error rate of accuracy to 15.3% in thousands of category classifications in The ImageNet Large Scale Visual Recognition Challenge (LSVRC). After this, the research [2-6] on deep learning techniques witnessed a development boom and a spurt of growth in the following years.

The field of medical image segmentation has also gradually made a qualitative breakthrough in this wave as a research boom, i.e., medical image segmentation[7-9], which aims to automatically extract highly expressive features related to locked lesions and extract lesions based on the features, has reached near-expert levels. Promoting automatic medical image analysis helps in treatment, prognosis, and medical research. However, at present, deep learning-based solutions still have a series of technical difficulties: (i) medical images contain less information than natural images, have lower contrast and poorer clarity, as well as include noise and artifacts that provide limited tissue descriptions; (ii) the structure of biological tissues is complex and varied, and the segmentation process needs to consider the mutual interference and interconnection of different tissues; (iii) the same class of targets in natural images generally have the same aspect ratio and similar morphology, while medical lesions usually have unconventional morphology and uncertain location relationship, which makes segmentation more difficult; (iv) Constrained by the imaging environment, operational norms, social ethics, and medical equipment and imaging standards, the number of accessible medical images is low, their quality is poor, and the images collected in different databases vary widely; (v) obtaining a more reliable segmentation performance on a limited and low-quality database requires a carefully designed better model structure. The above factors result in deep learning models that perform well in the natural image domain, but not as well in medical imaging tasks, especially in areas such as segmentation and generation of pixel-level prediction requirements, which pose challenges for future medical-aided diagnosis deployments and applications[7, 8, 10, 11].

Recent studies [12, 13] have shown that attention mechanisms are blossoming in both vision and language processing domains, constantly breaking through bottlenecks in model performance. The more intuitive and complex attention mechanisms are beginning to show explosive growth, which suggests that the mechanism has a positive guiding role for semantic understanding in machine vision and other domains. It is more common to pay attention to important information in the channel dimension and the space dimension of the feature dimension, i.e., SE Layer [3] and GE Net [14], when the self-attention mechanism [15, 16] in natural language processing (NLP) is transferred to form modules led by NL [17] and Vision Transformer [18], long-range point-to-point similarity attention has been widely proved to improve module performance since similar images usually have similar representations.

Most scholars [19, 20] are combining existing networks with an attention mechanism to refresh the new metrics of the task by exploiting the global focus property of the attention mechanism, adding attention to the model has proved to be an excellent mechanism, and it has become a trend to combine medical image semantic segmentation networks with an attention mechanism. At present, there is also a lot of work attempting to execute task-adaptation adjustments based on

the improved attention module, which fully demonstrates the flexibility and diversity of design and usually achieves stable metrics and higher results. In summary, the contribution of this paper has three main points:

1. The technical challenges in medical image segmentation are summarized, and essential guidelines are provided for solving these challenges through attention mechanisms.
2. The basic types of attention in recent years are collated and their respective advantages and disadvantages are presented to provide researchers with references for in-depth research and improvement of attention mechanisms.
3. Combining the current attention applications in medical image segmentation, we summarize the scenarios of each attention application and analyze the significance of future research in this field.

2. Application of Attention Mechanisms

2.1. Description

In neural networks, relying on convolutional operations in the current layer has a limited ability to modulate features locally, which may affect the accuracy of recognition [4, 18]. However, convolution operation cannot take into account the distribution of features beyond the current location and the interactions between features, and it is necessary to keep stacking these structures at a deep level to obtain a suitable long-range perception. The attention mechanism is essentially the process of focusing on a set of regions based on a more comprehensive perception of the environment than the model through comprehensive contextual information or given external conditions [21-23]. The mechanism is widely used to increase the weight of key location features while suppressing noise and interference from other locations. Currently, attention mechanisms can be divided into soft attention and hard attention according to the nature of differential derivability, the former of which is generally self-learning a set of weights that are back-propagated according to the gradient and adaptively updated with parameters. The latter enhances the attention of the model to the region of interest by giving a set of masks for region exclusion.

The section headings are in boldface capital and lowercase letters. Second level headings are typed as part of the succeeding paragraph (like the subsection heading of this paragraph). All manuscripts must be in English, also the table and figure texts, otherwise we cannot publish your paper. Please keep a second copy of your manuscript in your office. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

2.2. Conventional Attention

Channel domain attention can help the model to weight the feature channels according to the signals of spatial aggregation, which is represented by the scheme SE Net [3]. As shown in Figure 1, it designs a separate branch outside the main branch for obtaining the global correlation of the feature map in the channel dimension and then assigns the corresponding channel weights to the main branch, which dynamically adjusts the feature maps between channels accordingly. SK-Net [24], on the other hand, considers that different convolutional kernel sizes have different effects on attention, so two branches of small size 3×3 and large size 5×5 are designed to assign different influence fractions, and thus have good effects on feature recalibration at different scales. To achieve a lightweight channel attention design while normalizing the channel order, the order module [22] considers the channel patterns in the vicinity of the current channel, and patterns have a sliding window implementation using 1D convolution with shared parameters. To achieve relationship capture between multi-distant variables, GPoS [25] performs 2nd-order pooling to construct relationships between point pairs, so that the generated channel adjustment scores can be generated based on higher-order correlations.

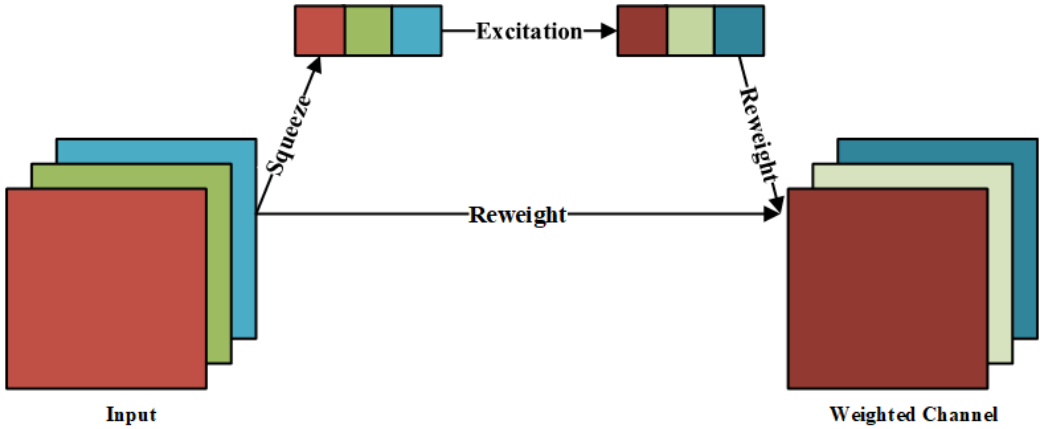

Figure 1 Operation of Squeeze-and-excitation channel attention

While channel attention has been shown to highlight features of beneficial channels, such attention typically ignores the positional relationships of spatial features, which are often critical to the need for spatial selectivity. Spatial information includes the location, shape and spatial relationships, and spatial structure of the target. Spatial domain attention then performs adjustments for feature values in the height and weight dimensions concerning spatial contextual information. The research [14] argues that local operators of CNN, e.g., convolution layer and pooling layer, can match some spatial information of an image, but may prevent the model from capturing remote feature interactions. Therefore, the Gather-Excite module, as shown in Figure 2, is proposed to perform feature map adjustment by first aggregating feature responses over a larger range and then merging the information and redistributing it to local features.

Experiments show that this mechanism can achieve gains consistent with increasing CNN depth at a low cost. The study [26] introduces a two-branch structure for information flow fusion-gathering attention branch and distributing attention branch, where the former self-learns a bi-directional relational mapping between locations that mitigates the local nature of convolution. The perceptual fields received by the network should be modulated in different environments. Coordinate Attention [27] decomposes channel attention into two parallel one-dimensional feature encodings by pooling features in different spatial directions, which preserves precise location information and captures the long-range dependencies.

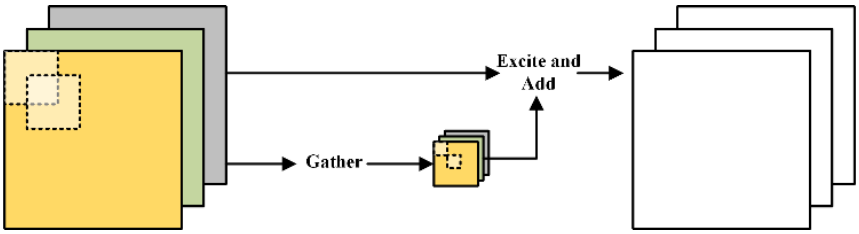

Figure 2 Operation of Gather-and-excite spatial attention

A more general technique is to focus on both the channel domain and the spatial domain for composite domain attention, such as CBAM [13] and BAM [12], which use series and parallel schemes to deal with the two dimensions separately, respectively, and the overall architectural differences between the two are shown in Figure 3.

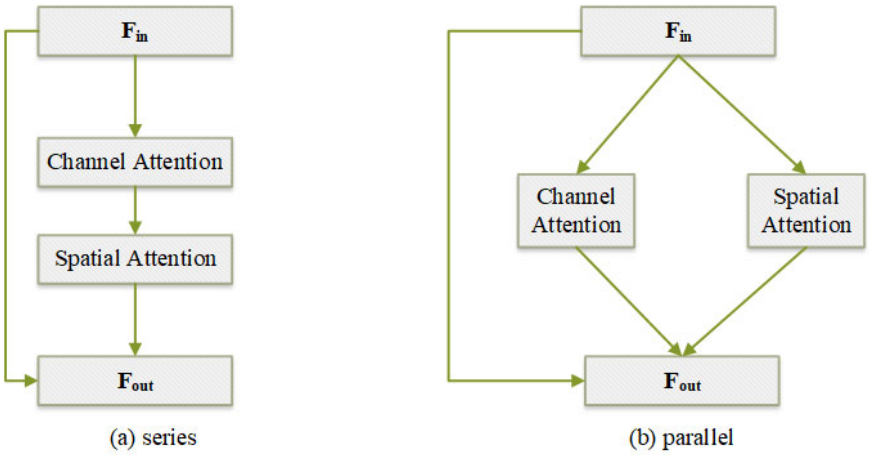

Figure 3 Main forms of channel-spatial compound attention

2.3. Multi-scale/Branching-based Attention

A diversified operation of feature extraction is the construction of parallel multi-branch structures that focus on different contexts, which usually provides a more comprehensive understanding of the network and enhances gradient conduction, SK-Net [24] can also be seen as the attention of two different convolutional kernel branches and therefore can enrich the representation of the model. Another option is to treat the constructed multi-branch results such as Inception [28, 29], Asymmetric Conv [30], or ASPP [31] modules discriminatively in different branches, with the output features of the current phase being generated by the best branch dominated by the best branch, supplemented by other branches for joint decision making. More meaningful work is the application of multi-branch attention to a convolutional extractor, which adjusts the four dimensions of the convolutional kernel to find the best pattern-matching state, allowing one operator to adapt to different feature scenarios. Figure 4 displays an example of the use of attention-guided multi-branch structures: Let l represent the current network layer, (a) this process facilitates the input feature map F_l to generate a set of attention maps \left\{A_l^n\in R^{N\times H\times W}\middle| n=1,2,\cdots,N\right\}, where N is equal to the total number of the multi-scale extraction process (b). Then, the mathematical logic of multi-scale/branching attention fraction calculation can be expressed as:

$$A_l = \tau\left(\mathrm{Conv}\left(\mathrm{ReLU}\left(\mathrm{Conv}_{1\times1}(F_l)\right)\right)\right) \tag{1}$$

In the above equation, the first convolution layer halves the number of channels and is used to reduce the computational effort; the second convolution aims to match the final number of channels to N. The function $\tau(\cdot)$ is an element-level Softmax function that limits the sum value of the feature at each of the same positions in channels dimension to 1. In this way, the distribution of inputs and outputs is normalized to some extent.

In the process of (b), there are a total of N different branches processing the contextual information, and the set of branch feature maps produced is $\{F_l^n | n = 1,2,\cdots,N\}$. In this process, the attention graph A_l^n for each channel uses the broadcast mechanism, which firstly performs element-wise multiplication with F_l^n separately, and then performs addition to obtain the input F_{l+1} at level $l+1$. The process can be described by the following equation:

$$F_{l+1} = \sum_{n=1}^{N} A_l^n \odot F_l^n \tag{2}$$

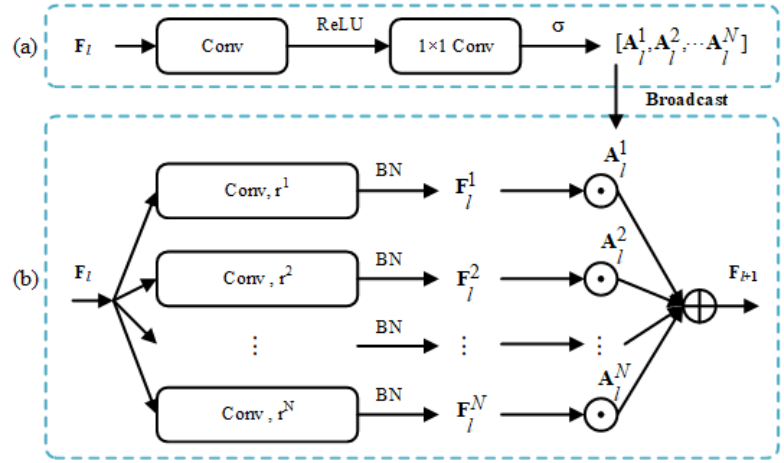

Figure 4 Attention on multiple scales/branches

2.4. Self-similarity Attention

The bottleneck of the above method is that the attention map of each pixel needs to be computed for the full map. Although the 1×1 convolution operation is utilized in NL to compress the dimensionality of the feature map, this global pixel-to-pixel (pixel-to-pixel) based modeling approach is computationally huge. The literature converts the original global interaction between self-attentive point-to-graphs to point-to-pyramid pooling features, thus reducing the computational complexity and the number of model parameters on Matmul and Softmax operations while having a multi-scale interaction structure. The literature [32] decomposes the global interaction in the spatial dimension and indirectly implements equivalent global interactions by alternating the execution of vertical and horizontal interactions. CC-Net [33] associates the pixel points of interest with the row and column pixels in which they are located one by one, and only two consecutive layers of cross-attention are stacked when the pixel points to be interacted with are not directly associated. The dimensional decomposition scheme reduces the computation of self-attention by 1~2 orders of magnitude and improves execution efficiency. Directly acquiring global connections is not an optimal solution and is prone to computational overload. The literature [34] rearranges the spatial locations and constructs a two-step scheme for long-range self-attention and short-range self-attention, and the relationship between any two locations can be captured by the tandem execution of both. Unlike the channel- and space-based mechanisms, inspired by convolutional locality, the literature [35] proposed a local content-based self-attentive mechanism, which makes the number of self-attentive parameters independent of the perceptual field size, and the number of operations is greatly reduced. In the literature [36], the Expectation-Maximization (EM) algorithm was introduced into the attentional graph to solve the computationally intensive problem of non-local operations by iteratively solving a compact set of basic features and modeling the interpixel connections in low-latitude manifolds. This scheme reduces intra-class variation while maintaining inter-class variation.

2.5. Other Attention

In addition, there are attention mechanisms designed by focusing on other perspectives, such as frequency domain attention, category attention, temporal attention, and recent works in this area include FcaNet [37], IAU-Net [38], and OCR-Net [39], respectively. Table 1 shows the classification and representative works of attentional mechanisms.

Table 1. Classification table of attentional mechanisms

class	type
Conventional attention	Channel-dimension attention, spatial-dimension attention, dual-dimension attention, etc.
Self-similarity attention	Unidimensional self-attention, multidimensional self-attention, hybrid self-attention, etc.
Other attention	Hard attention, frequency domain attention, category attention, etc.

3. Literature References

The main imaging principle of medical imaging is to scan a specific area with sophisticated instruments and collect feedback signals that are converted into a visually friendly structured form. Medical images are often low-contrast and inhomogeneous-looking images, making it difficult for a well-designed model to perform theoretically if the segmentation model cannot focus on key information. Attention mechanisms can guide downstream tasks to grasp discernible expected features, filter and suppress irrelevant information, enhance the strength of target features, and improve network decision accuracy. A well-designed attention module can replace higher accuracy results with less parameter cost, a feature that is important in the field of medical segmentation, where precise and efficient localization of ROI regions such as brain tumors is the most essential requirement in clinical application scenarios such as preoperative tumor confirmation, tumor location tracking, and condition analysis. In addition, since schemes to enhance feature perception often make the network structure complex and diverse, selecting key information and discarding redundant information is fundamental to focus on core region features and optimize model prediction efficiency.

3.1. Application of Conventional Attention

Currently, many medical image segmentation-based models simulate the focusing mechanism of biological vision by pairing attention modules with conventional attention mechanisms for feature reweighting based on feature contextual interactions to further enhance the evaluation metrics. Attention U-Net [1] pioneered the concept of

attention gate (AG), which guides jump connections and corrects the strength of features with deep high-level semantics, and AG implicitly emphasizes task-relevant transfer while reducing task-irrelevant interference. The final experiments show that AG can improve the predictive power and guarantee the efficiency of U-Net. Li et al [2] proposed the attention nested U-Net (ANU-Net) based segmentation network for organ cancer segmentation, which introduced a redesigned dense jump connection and attention mechanism based on Attention U-Net, and designed a new hybrid loss function, which was used in the liver tumor segmentation dataset LiTS (liver tumor segmentation), and healthy abdominal organ segmentation CHAOS (combined healthy abdominal organ segmentation) datasets both achieved top segmentation results. Some schemes use multiple branches for weighting different attention, so the model can have more expressive pooling in the context and achieve segmentation performance improvements. Since deep supervision was found to improve performance, imposing attention on multiple decoders and using their results as a common output is a good scheme [88] that can enhance prediction adaptation to different scales, e.g., CA-Net [3] applies channel- and space-level attention to the output of multiple decoding stages before performing prediction. In retinal vascular segmentation, due to the complexity and inconspicuousness of blood vessels, it is necessary to combine multiple scales for joint decision-making, and the scale multiplexing mechanism of Res2Net can establish a connection between small and large scales; not only that, SA-Net [4] applies the attention mechanism to multiple scales to balance the importance of multiple scales, so good segmentation results are obtained. DDANet [5] designed two decoders with attention residual blocks and staggered fusion in the decoder. The attention mechanism successfully helped DDANet to obtain a good polyp outlining ability and finally achieved a dice score of 85.76%. There are also many excellent types of research [6, 7] that exploit the attention application on the medical image segmentation domain.

3.2. Application of Self-Attention

Recently, with the success of non-local blocks and Transformers in the field of vision, combining medical image segmentation with self-attentive mechanisms is also a direction worth exploring.

Many excellent models of self-attentive mechanisms enable models with more complete perceptual capabilities, and the literature [45] introduced scalable resolution NL blocks as well as improved residual blocks in the codec structure, which are not only applicable in each stage but also can be added to the downsampling and upsampling layers, ultimately achieving several metrics improvements in the infant brain segmentation task. M3-Net [46] constructed two paths for encoding stage-based interactions to perform pancreatic segmentation with biphasic three-views, and each of the encoding terminals was inserted with an NL-based local attention and depth attention, and this scheme achieved the highest Dice coefficients of 91.19% and 86.34% in the internal and public pancreatic datasets, respectively. CaraNet [47] considered the lack of global similarity interactions in RA and therefore added axial self-attention to the RA module to achieve the highest segmentation accuracy. The literature [48] used resnet as a backbone to extract local structural information based on convolutional networks, and stacked Transformer modules at the end to extract similarity-dependent information based on a self-attention mechanism, demonstrating that Transformer-based networks can also be successful in medical image segmentation. u-Net Transformer [49] introduced a multi-headed attention mechanism at the deep level based on attention U-Net for global understanding between point-to-semantics and replaced the attention gate module with inserted crossover self-attention in the decoding stage to filter irrelevant information based on high-level semantics and similarity scores of jump connections. Recent works such as Swin U-Net [9] and Swin U-Netr [50] have obtained segmentation performance that is not weaker than pure CNNs, which are based on the practical demonstration of Swin Transformer [51] models in 2D and 3D medical segmentation, and these schemes decouple the feature extraction process into inter-region extraction patterns and intra-region extraction patterns Implemented to indirectly construct inter-region information transfer by offsetting regions, this decomposition mode can enhance the intra-region attention and extract more reasonable features

In summary, combining attention mechanisms with feature perception techniques results in highly accurate and robust task performance, Figure 5 displays several medical image applications. Common segmentation schemes for attention-guided perception include jump-connection-based attention-guiding, multiscale attention-guiding, multidimensional interaction-based attention-guiding, and mask-based attention-guiding. Correction of feature strength with attention based on multiple modalities in different contexts not only allows for flexible and reasonable features, but also the model can learn to change its output or behavior to be placed in the best representation. Applying the attention-based segmentation scheme to brain tumor segmentation can assist the model to perceive multiple tumor area locations and size states, focus on discriminative features of regular brain tissue and abnormal tissue, and can effectively avoid over- and under-segmentation states, and improve tumor detection rates.

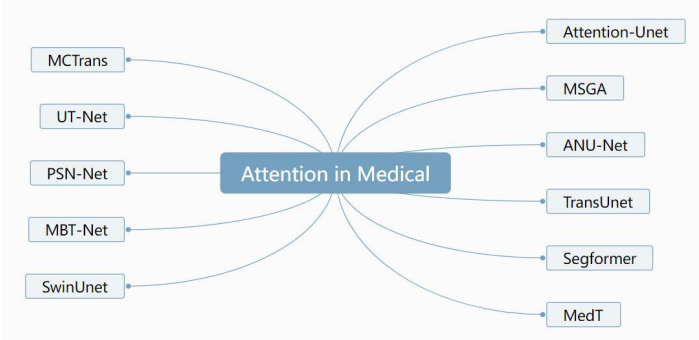

Figure 5 Common Attention Schemes for Medical Image Applications

4. Conclusion

Appropriate employment of attention mechanisms in medical imagery can assist baseline focus on medical ROI-related prospects and features. This enables network optimization to focus near the target and perform fine-tuning, and ultimately benefit from the model's evaluation metrics. In conclusion, the attention mechanism plays an important role in medical image segmentation, which can help the model accurately segment tissues and organs in medical images, thus improving the efficiency and accuracy of medical image analysis. In practice, a slight improvement can reduce the risk of segmentation errors and misdiagnosis and then greatly improves the cure rate, so it has a strong research value.

In summary, the practical significance of attention-based medical image analysis technology research mainly includes the following aspects:

1. Diagnosis and treatment: assist physicians in more accurately determining the location, size, shape, and type of tumors. For example, the tumor area obtained by segmentation can be used for preoperative planning and intraoperative localization, helping to precisely remove the tumor and minimizing damage to normal brain tissue.

2. Research and education: it can be utilized to compare different lesion types, grades, sizes, and locations, while it points in an investigating direction of the mechanisms of tumorigenesis, growth, and spread for medical education. In practice, t physicians' knowledge and ability to diagnose can be improved rely on this technique.

3. Improving medical efficiency: Manual examinations require specialized knowledge and are less efficient, and supplementing manual reading with low-cost machine reading for disease screening and diagnosis can reduce physician stress and improve efficiency.

4. Balancing regional resources: The varying levels of physicians in different regions make it difficult to balance and meet the growing demand for medical care, and automated medical analysis has a positive impact on balancing regional resources and alleviating social conflicts

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[5] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492-1500.

[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, 2014: Springer, pp. 818-833.

[7] M. Mubashar, H. Ali, C. Grönlund, and S. Azmat, "R2U++: a multiscale recurrent residual U-Net with dense skip connections for medical image segmentation," Neural Computing and Applications, vol. 34, no. 20, pp. 17723-17739, 2022.

[8] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, 2022: Springer, pp. 23-33.

[9] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, 2023: Springer, pp. 205-218.

[10] S. Xun et al., "Generative adversarial networks in medical image segmentation: a review," Computers in Biology and Medicine, vol. 140, p. 105063, 2022.

[11] V. Thambawita et al., "SinGAN-Seg: Synthetic training data generation for medical image segmentation," PloS one, vol. 17, no. 5, p. e0267976, 2022.

[12] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," arXiv preprint arXiv:1807.06514, 2018.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.

[14] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," Advances in neural information processing systems, vol. 31, 2018.

[15] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[16] F. Sun et al., "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441-1450.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794-7803.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[19] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, and D. Yang, "EANet: Iterative edge attention network for medical image segmentation," Pattern Recognition, vol. 127, p. 108636, 2022.

[20] J. Cheng et al., "ResGANet: Residual group attention network for medical image classification and segmentation," Medical Image Analysis, vol. 76, p. 102313, 2022.

[21] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11794-11803.

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534-11542.

[23] H. Zhang et al., "Context encoding for semantic segmentation," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7151-7160.

[24] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510-519.

[25] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2019, pp. 3024-3033.

[26] H. Zhao et al., "Psanet: Point-wise spatial attention network for scene parsing," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 267-283.

[27] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13713-13722.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[29] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.

[30] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1911-1920.

[31] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[32] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," arXiv preprint arXiv:1912.12180, 2019.

[33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 603-612.

[34] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," arXiv preprint arXiv:1907.12273, 2019.

[35] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," Advances in neural information processing systems, vol. 32, 2019.

[36] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9167-9176.

[37] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 783-792.

[38] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "IAUnet: Global context-aware feature learning for person reidentification," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 4460-4474, 2020.

[39] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, 2020: Springer, pp. 173-190.

[40] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[41] C. Li et al., "ANU-Net: Attention-based Nested U-Net to exploit full resolution features for medical image segmentation," Computers & Graphics, vol. 90, pp. 11-20, 2020.

[42] R. Gu et al., "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," IEEE transactions on medical imaging, vol. 40, no. 2, pp. 699-711, 2020.

[43] J. Hu, H. Wang, J. Wang, Y. Wang, F. He, and J. Zhang, "SA-Net: A scale-attention network for medical image segmentation," PloS one, vol. 16, no. 4, p. e0247388, 2021.

[44] N. K. Tomar et al., "DDANet: Dual decoder attention network for automatic polyp segmentation," in Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII, 2021: Springer, pp. 307-314.

[45] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local u-nets for biomedical image segmentation," in Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, no. 04, pp. 6315-6322.

[46] T. Qu et al., "M3Net: A multi-scale multi-view framework for multi-phase pancreas segmentation based on cross-phase non-local attention," Medical image analysis, vol. 75, p. 102232, 2022.

[47] A. Lou, S. Guan, and M. Loew, "CaraNet: context axial reverse attention network for segmentation of small medical objects," Journal of Medical Imaging, vol. 10, no. 1, p. 014005, 2023.

[48] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[49] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12, 2021: Springer, pp. 267-276.

[50] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I, 2022: Springer, pp. 272-284.

[51] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012-10022.