A Survey of Medical Image Segmentation Based on The Pipeline of U-Net

Heyu Huang, Lianhe Yang

***Abstract*— Deep learning methods have been successfully applied to implement medical image segmentation. It employs convolutional neural networks (CNN) to compute distinctive image features from a defined pixel-wise objective function. However, it has become a powerful tool for computer-aided diagnosis (CAD), where multi-scale fine-grained learning of segmentation targets and global contextual interdependencies have been under constant exploration and progress, they are still the main reasons for the non-robust and unrealistic segmentation of medical images. After extensive research work, we propose in this paper five new practical and advanced methods from different perspectives, aiming to alleviate and counteract the problems in current state-of-the-art methods. They are developed using U-Net as a baseline, focusing around a multi-scale approach guided by attention flow thus obtaining fine-grained features and global contextual information. The methods in this paper make references to help neural network models improve the accuracy of segmentation results, and try to promote the development of the field of wise medical care with the goal of practical applications, and work towards the co-creation of artificial intelligence to lead medical progress.**

***Keywords*— Medical image segmentation, CNN, Attention mechanism, Multi-scale features**

I. Introduction

A. Medical image segmentation based on deep learning

Medical imaging plays a vital role in medical clinical treatment and diagnosis. They can provide non-destructive anatomical and functional information about diseases or abnormalities in vivo through medical imaging techniques such as X-ray, CT and MRI. Among them, medical image segmentation [1, 2] is a key step in pathological assessment, treatment planning and disease progression monitoring, and it is a process of segmenting images into multiple meaningful or regions of interest that can be used for better clinical analysis or visualization of clinical targets. Figure 1 shows the medical images obtained by different imaging methods. However, there is a need to obtain accurate and robust image segmentation results in clinical routines, which are highly dependent on the imaging modality and target (anatomical or focal region). Therefore, further exploration in medical image segmentation methods is needed. Image processing techniques as a powerful tool in Computer-Aided Diagnosis (CAD) systems [3] can facilitate CAD systems to better track diseases and thus better treat them, and to save a lot of time and cost in manual processing and reduce errors arising from subjective variability. However, the need for such an accurate and reliable automatic segmentation method that alleviates the workload of medical specialists such as radiologists is high and remains a challenging task. First, due to the limitations of imaging equipment and the imaging environment, medical images are generated with varying degrees of interference from artifacts, speckles, and other noise, making it extraordinarily difficult to obtain more useful and important discriminative information from image segmentation. Secondly, medical images are different from other images due to the specificity of their own properties, in which the segmentation targets have complex geometric shapes and low contrast, i.e., the foreground and background regions are confused with each other, and their structures vary from person to person and their positions are flexible, which is very likely to lead to incomplete segmentation or even misclassification. It increases the difficulty of constructing a model with respect to a priori shape constraints, which can lead to unrealistic and non-robust segmentation results.

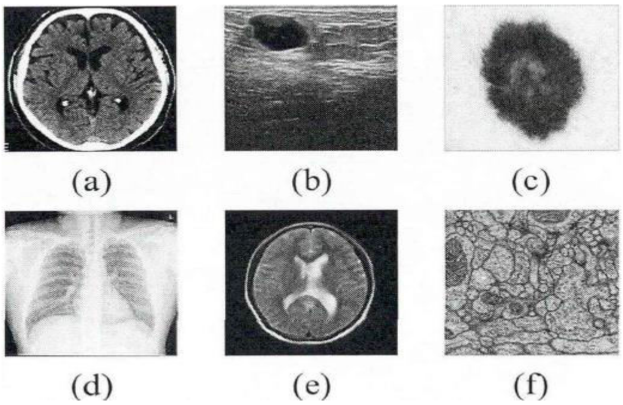

Fig. 1. Images obtained by different medical imaging devices. (a) brain images obtained by computed tomography (CT), (b) breast lesions obtained by ultrasound equipment, (c) skin lesions obtained by dermatoscopy, (d) chest films obtained by X-ray, (e) magnetic resonance imaging (MRI) of the brain, (f) cellular tissue structures obtained by electron microscopy

In recent years, convolutional neural networks (CNNs) [4] have achieved state-of-the-art (SOTA) performance in a wide range of vision tasks due to their powerful nonlinear and hierarchical feature extraction capabilities. These excellent deep neural network models have been applied and dominated in the field of medical image segmentation, achieving excellent performance and results, which makes automatic segmentation methods based on deep learning techniques a reliable alternative solution. In this context, the proposed full convolutional neural network (FCN) [5] with

Manuscript received February 10, 2022

Heyu Huang, The faculty of Computer Science, Tiangong University, Tianjin, China
Lianhe Yang, The faculty of Computer Science, Tiangong University, Tianjin, China

codec architecture, as shown in Figure 2, and then the proposed U-shape Net (U-Net) [6] framework, after being widely used in medical image segmentation and showing faster and more accurate performance, in fact, U-Net and variants of the model [7, 8] have become the gold standard in medical image segmentation. These architectures typically consist of a systolic path, which compresses the input image into a set of high-level features, and an extended path, which uses high-level features to reconstruct pixel segmentation masks at single or multiple upsampling steps, which are then connected using low-level and high-level features with different semantic information. In addition, multi-scale jump connections [9, 10], expanded convolution [11, 12], and dense convolution blocks [13, 14] are embedded in the U-Net framework and used to further improve the segmentation accuracy.

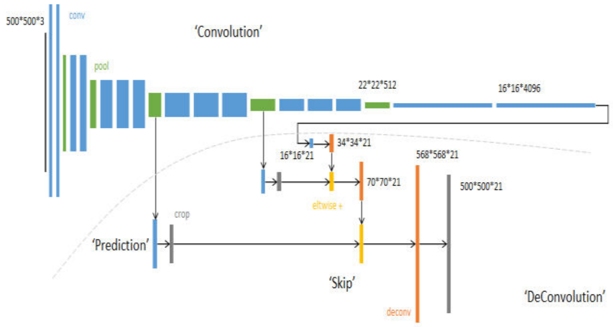

Fig. 2. Structure diagram of Full Convolutional Neural Network (FCN)

B. *Multi-scale feature aggregation*

On the basis of the above exploration, it is desirable to consider and learn from multi-scale fusion strategies [15, 16] for semantic segmentation tasks, aiming to achieve full integration of detailed and semantic information in medical images more effectively and obtain a strong feature representation with both spatial and semantic, which is beneficial to capture targets of varying structural sizes and acquire discriminative features in medical images. However, how to perform the integration and achieve fruitful results needs further exploration. As shown in Figure 3, the use of Pyramid Pooling Module (PPM) is proposed in PSPNet [17, 18] for modeling multiple scales of the same target. And DeepLab series [19, 20] use the operation of Astrous Spatial Pyramid Pooling (ASPP) to indirectly obtain multiple scales for feature aggregation.ICNet [21] sets the input as a multi-scale image and uses a cascade method to aggregate features for efficiency.

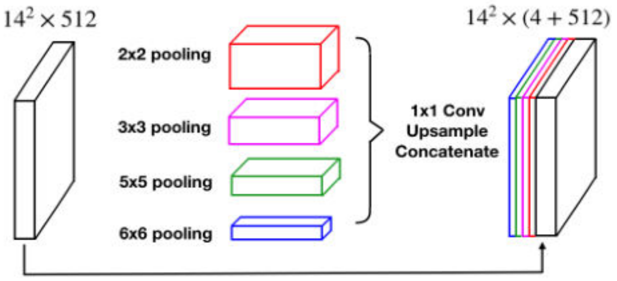

Fig. 3. Structure diagram of a conventional multi-scale feature aggregation module

However, as stated in [22, 23], the multiscale fusion operation that only performs stepwise aggregation for adjacent scales in a single direction will continuously increase the information loss of semantic features during propagation, resulting in ineffective propagation of semantics from deep to shallow layers. This is suboptimal for medical image segmentation, which not only leads to the loss of more discriminative features in segmentation targets or lesion regions that are not rich enough in semantic information, but also causes the ineffectiveness of multiscale stacking and increases the network complexity, which is even worse for medical images that already have a small number of samples and are highly susceptible to overfitting.

C. *Attention mechanism*

Besides, the proposed attention mechanism [24, 25], which aims to focus on the most relevant information and suppress the behavior of irrelevant parts in the feature map, is extremely friendly for medical image segmentation by highlighting key information while suppressing the interference of noise and similar categories. The most basic model of channel attention mechanism is shown in Figure 4. Therefore, to further improve segmentation performance, attention mechanism has received extensive attention in the field of medical image segmentation by effectively highlighting useful parts and suppressing redundant information in images, and many studies have also attempted to embed attention modules into network architectures to improve performance. Hu et al [26] proposed a squeeze and stimulate (SE) module as a novel attention module that focuses on channel relations and dynamically performs channel feature realignment to enhance feature representation. Yu et al [27] input semantically stronger deep features into the SE-like attention module to provide high-level category information that helps to accurately recover details during the upsampling phase of the image segmentation process. Unlike the use of attention mechanisms to re-estimate important information, other studies have focused on capturing the long-term dependence of global context. Zhang et al [28] extend the nonlocal module using prior distributions and construct the set of nonlocal operations with weights to pursue better segmentation performance. Fu et al [29] propose a dual attention module consisting of spatial attention and channel attention to implement semantic segmentation, where the dual attention module is similar to a nonlocal operation [30, 31]. Although these works make use of attentional mechanisms to highlight important features and suppress background interference to some extent, the attentional results of these approaches are generated independently of the corresponding feature levels. Tending to the idea in [32, 33], for segmentation targets with smaller structures, the responses regarding attention do not emphasize enough the local or overall contour details of the small target and are less useful and interpretable for a comprehensive understanding of the use of attention in medical image segmentation.

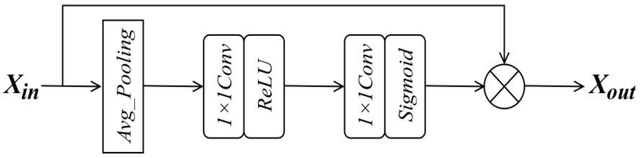

Fig. 4. Structural diagram of the channel attention mechanism (CA)

D. *Existing challenges*

In summary, although today's automatic medical image segmentation methods based on deep learning technology have become mature and stable, I personally believe that the task has more room for improvement and obvious problems need to continue to explore, summarized as follows: (1) medical image properties and imaging conditions and other objective reasons, resulting in medical images by artifacts and other noise interference is strong, and the background or similarity of false positive category contrast is low, and the segmentation target structure is diverse, the location is uncertain, the regional boundary is blurred, the category imbalance problem is serious; (2) medical image data collection is difficult and costly, the limited number of samples limits the learning ability of neural networks, and is very easy to cause overfitting; (3) the current superior performance of methods such as U-Net-based improved model whose scale is large, and consume a lot of computing resources and memory; (4) the current superior performance of methods, such as U-Net-based improved model of its large scale, often requiring millions of parameters, consuming a large number of computing resources and memory, hindering the successful implementation of most excellent medical image segmentation methods into embedded devices; (5) the lack or ineffectiveness of the acquisition of global contextual information is the cause of incomplete segmentation, segmentation results are not The key and stubborn reason for the unrealistic segmentation results.

II. Evaluation Method

Medical image segmentation belongs to one of the semantic segmentation tasks of vision, and due to its special image nature, a binary classification operation is usually performed for the segmentation target. Therefore, the evaluation of segmentation performance is particularly critical. In order to better evaluate the segmentation methods, i.e., the individual pixels should be identified and labeled, the majority of the work uses The Dice coefficient (DSC) and the Intersection over Union (IoU), the two most commonly used metrics. As Equation 1 demonstrates the mathematical DSC solution method. The DSC used is for comparing the similarity between the resulting segmentation results and the original ground truth, while the IoU is for comparing the overlap between the output mask of the segmented target and the original ground truth mask. The mean Intersection over Union (mIoU) then calculates the IoU for each semantic class of the image and finds the average of all classes. there is a correlation between DSC and mIoU. Therefore, many works provide a better understanding of the results by calculating these two metrics in order to provide a comprehensive analysis of the results.

$$DSC = \frac{2\left|y \cap \hat{y}\right|}{\left|y\right| + \left|\hat{y}\right|} = \frac{2TP}{2TP + FP + FN} \tag{1}$$

where the term $\cap$ denotes the overlap between the segmented region $\hat{y}$ and the ground truth y , and $\left|\ \right|$ denotes the number of voxels belonging to each ROI.

III. Research Idea And Methodology

In this paper, we propose the following five aspects of research and experiments to overcome the above difficulties one by one, based on the problems and challenges of the deep neural network model for automatic segmentation of medical images.

A. *Lightweight Network*

Lightweight refers to the design of a module by parallelizing or sharing computational blocks in a way that reasonably stacks and places the lightweight backbone of the composed model, taking the AHSP modules in Inf-Net [34] and MiniSeg [35] as references and targets, and striving to find a balance between performance and computational volume. Directly, it is also possible to draw on the real-time semantic segmentation models of the last two years such as ESP-NetV2 and MobileNetV3 as the backbone of their own models directly for solving the problem of overfitting or training gradient dispersion and real-time landing easily caused by the huge number of parameters of existing methods and the small relative number of samples.

B. *Comprehensive Attention Stream Method*

In the overall segmentation network, we embed comprehensive attention methods with different focus to correctly deal with the properties and meanings of features at different levels or stages (encoding and decoding features, spatial detail features and semantic features), improve the use of redundant information caused by the simple stacking and connecting features without differentiation in the previous methods, and correctly control the different important properties of different features and their importance. It is well known that shallow features have better spatial structure information, while deeper features have more comprehensive and richer semantic information, so how to use the attention flow to intersperse the different nature of the interlayer structure, and how different attention methods can complement each other for different features, can be extended from this direction.

C. *Modifying Constraints Such As Loss Function*

As mentioned above, the targets to be segmented in medical images are special in nature, and the structure is variable and varies from person to person. For example, in brain tumor segmentation, the diffusion region to be segmented is closer to circular structure and arc-shaped structure, and in lung inflammation segmentation, the lesion region is similar to columnar or filamentary shape. It is possible to add a constraint regular term for the shape of the target to be segmented to the loss function used for constraint (e.g., Bce Loss, etc.), such as curved shape corresponding to convex function, so that the segmentation shape becomes a new supervision to improve the classification error or missed segmentation caused by low contrast and blurred boundary in segmentation. Alternatively, instead of using L1 or L2 or Dice at the pixel-wise level as a similarity measure, the boundary region of the segmentation target is extracted first for segmentation, and then Cos similarity is measured with GT to achieve the purpose of constraining the boundary, etc.

D. *Adding Few-shot Learning*

The Few-shot learning approach is used to solve the problem of limited sample data for medical images, i.e., to achieve excellent segmentation performance using limited data. Based on our thinking, the reason for not choosing weakly supervised and unsupervised methods is that, firstly, for traditional medical images (MRI, CT and ultrasound, etc.) unlike semantic segmentation in scene images, the categories classified are less (generally binary), plus their categories are severely unbalanced, so when methods such as weakly supervised confirm the categories by calculating probability maps, it is extremely easy to incorrectly calculate false positive categories as segmentation targets. Secondly, when the category location is determined by the heat map obtained through the CAM method, although it is difficult to obtain discriminative features of medical images, the semantic information carried by them is not rich (e.g., tumor lesions, inflammatory regions and polyps, etc.), so the CAM algorithm calculates the heat map to determine the target location based on the semantic information, which may also easily cause the false positive category to be misclassified and interfere with the segmentation process. In summary, there are good reasons to consider weakly supervised and unsupervised methods as suboptimal in this aspect of the solution.

Therefore, a Few-shot learning approach is envisioned to accomplish medical image segmentation. The initial idea is to divide the segmentation network into two steps, with the first step performing Few-shot learning, which makes the network perform a coarse identification of the segmented targets, avoiding as much as possible the interference of false positive categories and boundary blurring problems at the beginning. It is possible to manually divide a small set of samples according to the typical category of the segmented target or the properties of medical images (e.g. MRI sequence, ultrasound image phase period, etc.) to form a k-way 1-shot or 2-shot, which only allows the network to identify or classify the segmented target and effectively determine the foreground and background; then the second step of segmentation is carried out, using the previous "classification" result as the basis. Then the second step of segmentation is performed, using the previous "classification" result as a priori, and then the segmentation is performed under supervised conditions, thus improving the segmentation performance even under the condition of limited data samples.

E. *Adopting Ensemble Learning*

Our idea is to join the super-resolution (SR) task to do branching and form a joint learning network to try to solve the medical image quality problem, i.e., enhance the image before segmentation, which makes the image enhanced and then do the segmentation with refinement, which helps the performance improvement. There is only one CAS paper on the idea of semantic segmentation combined with super-resolution and segmentation for cityscape images [36], which has not yet appeared in the field of medical images, so we think it is more innovative and feasible.

First, medical images such as CT, MRI, etc., have poor contrast while the image is blurred with unclear boundaries, so the image quality can be improved with the help of super-resolution task to generate higher resolution images, highlighting the shape structure and location of segmentation targets to make them more delicate. Secondly, super-resolution is better for recovering the spatial and structural information of the target in the image when the scale and resolution are recovered by sub-pixel convolution, and it can guide and facilitate the segmentation process by using the super-resolution features in the decoding stage to guide the segmentation features and promote the reduction of the difference between the two features by a method similar to the "parallax" supplement. Finally, with regard to computational considerations, the two tasks can share the encoder and use a lightweight encoding process to extract important discriminative information, and then divide into two tasks in the decoding process, which can add a perceptual loss between the two tasks to constrain the segmentation process more, and also allow for information transfer and feature complementation.

IV. Conclusion

In this paper, we propose innovative and practical ideas and approaches in five areas based on current problems and challenges in implementing medical image segmentation based on deep learning techniques. These approaches aim to address or alleviate the memory-consuming and computationally expensive problems caused by the large number of model frame parameters, the lack of global contextual information guided by attention streams, the problem of non-robust and unrealistic segmentation performance due to image quality and attributes, the problem of difficult access to medical image data and the problem of strong constraints on segmentation targets in medical image segmentation, respectively. This paper not only analyses the potential and significance of medical image processing and dissects the problems that exist in current learning methods, but also analyses and proposes model designs with new popular methods and thinking. It is hoped that it can provide ideas and references for the majority of vision scholars in the direction of deep learning.

References

[1] Z. Gu et al., "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," IEEE Transactions on Medical Imaging, pp. 1-1, 2019.

[2] H. Huang et al., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1055-1059, 2020.

[3] H. Fujita, "AI-based computer-aided diagnosis (AI-CAD): the latest review to read first," Radiological Physics and Technology, vol. 13, pp. 6-19, 2020.

[4] G. J. S. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[5] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 640-651, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015.

[7] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," IEEE Transactions on Medical Imaging, vol. 39, pp. 1856-1867, 2020.

[8] C. Li et al., "Attention Unet++: A Nested Attention-Aware U-Net for Liver CT Image Segmentation," 2020 IEEE International Conference on Image Processing, pp. 345-349, 2020.

[9] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation," Neural networks : the official journal of the International Neural Network Society, vol. 121, pp. 74-87, 2020.

[10] S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A New Multi-Scale Backbone Architecture," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, pp. 652-662, 2021.

[11] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature methods, 2020.

[12] D. Liu, H. Zhang, M. Zhao, X. Yu, S. Yao, and W. Zhou, "Brain Tumor Segmention Based on Dilated Convolution Refine Networks," 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications, pp. 113-120, 2018.

[13] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes," IEEE Transactions on Medical Imaging, vol. 37, pp. 2663-2674, 2018.

[14] Z. Yang, P. Xu, Y. Yang, and B. Bao, "A Densely Connected Network Based on U-Net for Medical Image Segmentation," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 17, pp. 1 - 14, 2021.

[15] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-Scale Interactive Network for Salient Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9410-9419, 2020.

[16] S. Gao, Q. Han, Z.-Y. Li, P. Peng, L. Wang, and M.-M. Cheng, "Global2Local: Efficient Structure Search for Video Action Segmentation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16800-16809, 2021.

[17] J. Zhou, M. Hao, D. Zhang, P. Zou, and W. Zhang, "Fusion PSPnet Image Segmentation Based Method for Multi-Focus Image Fusion," IEEE Photonics Journal, vol. 11, pp. 1-12, 2019.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 6230-6239, 2017.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 834-848, 2018.

[20] H. Wang, Y. Zhu, B. Green, H. Adam, A. L. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation," 2020.

[21] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for Real-Time Semantic Segmentation on High-Resolution Images," ArXiv, vol. abs/1704.08545, 2018.

[22] Q. Zhang et al., "Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images," IEEE Transactions on Image Processing, vol. 30, pp. 1305-1317, 2021.

[23] L. Wang, J. Zhang, Y. Wang, H. Lu, and X. Ruan, "CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss," 2020.

[24] Y. Liu, X.-Y. Zhang, J. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically Attentive Multi-Scale Network for Lightweight Salient Object Detection," IEEE Transactions on Image Processing, vol. 30, pp. 3804-3814, 2021.

[25] A. Vaswani et al., "Attention is All you Need," ArXiv, vol. abs/1706.03762, 2017.

[26] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, pp. 2011-2023, 2020.

[27] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a Discriminative Feature Network for Semantic Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1857-1866, 2018.

[28] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-Occurrent Features in Semantic Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 548-557, 2019.

[29] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual Attention Network for Scene Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3141-3149, 2019.

[30] X. Wang, R. B. Girshick, A. K. Gupta, and K. He, "Non-local Neural Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794-7803, 2018.

[31] Y. Cao, J. Xu, S. C.-F. Lin, F. Wei, and H. Hu, "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond," 2019 IEEE/CVF International Conference on Computer Vision Workshop, pp. 1971-1980, 2019.

[32] R. Gu et al., "CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation," IEEE Transactions on Medical Imaging, vol. 40, pp. 699-711, 2021.

[33] C. Li et al., "ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection," IEEE Transactions on Cybernetics, vol. 51, pp. 88-100, 2021.

[34] D.-P. Fan et al., "Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images," IEEE Transactions on Medical Imaging, vol. 39, pp. 2626-2637, 2020.

[35] Y. Qiu, Y. Liu, and J. Xu, "MiniSeg: An Extremely Minimum Network for Efficient COVID-19 Segmentation," 2021.

[36] L. x. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual Super-Resolution Learning for Semantic Segmentation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3773-3782, 2020.