GST: A Brain-Inspired Graph Signal Transmitter for Biomedical Image Segmentation

Caiqing Jian [1], Yongbin Qin[1*], Lihui Wang[1], Hui Xia [2], Yuemin Zhu[3]

[1] *Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis of Guizhou Province,*
State Key Laboratory of Public Big Data,
College of Computer Science and Technology, Guizhou University, Guiyang, China.

[2] *Department of Thoracic-cardio Surgery, Fourth Medical Center, PLA General Hospital, Beijing 100048, China.*
[3] *Univ Lyon, INSA Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France.*

Email: gs.cqjian21@gzu.edu.cn, Corresponding Author: wlh1984@gmail.com,
11734048@zju.edu.cn, xiahui304@163.com, zhu@creatis.insa-lyon.fr

2022 16th IEEE International Conference on Signal Processing (ICSP) | 978-1-6654-6056-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSP56322.2022.9965289

Abstract—**Automatic segmentation of biomedical image is of great significance for computer-aided diagnosis. Different from the existing segmentation models based on convolutional neural network or attention mechanism with pixelwise prediction mode, this paper proposes a new Graph Signal Transmitter (GST) which regresses patch probability directly from the evolved semantic node and edge features. The images pass through an encoder to derive the semantic features, each voxel of which is taken as a node, and the corresponding feature vector along the channel dimension is taken as the node feature of the initial graph layer, and the relative positions between different nodes are considered as the edge features. With the feedforward process, the node and edge features of the following layers are evolved by fusing the features of source nodes, destination nodes and edges. The node features of the last graph layers are directly mapped back to the patch in semantic space with fold module and generate segmentation map. In the semantic and instance segmentation tasks for gland dataset, comparative experiments show that the proposed model outperforms the state-of-the-art (SOTA) methods, and ablation experiments demonstrate the effectiveness of GST.**

Keywords—graph, brain-inspired, signal transmission, biomedical image, segmentation, deep learning.

I. INTRODUCTION

Segmentation is a challenging problem in medical image processing, it is usually implemented manually by the experienced clinicians, which is not only time-consuming and labor-intensive, but also introduces the problem of inconsistent labeling standards. In recent years, deep learning has achieved state-of-the-art performance on many applications in medical image segmentation, which is expected to solve the above problems.

As a milestone for semantic segmentation, the fully convolutional network (FCN) [2] enables CNN to be naturally adapted to pixelwise dense prediction tasks. Based on FCN, several improved models have been proposed. For instance, the models of DeepLab family [3-5] consider contextual information in the semantic segmentation through dilated convolution, PSPNet [6] combines multi-scale features to improve model performance, UNet and its variants [7-9] introduce skip connections between encoder and decoder to promote the segmentation details and have become benchmark models for medical image segmentation.

With the success of Transformer [10], attention mechanisms have gradually been applied into vision tasks. For example, Non-local UNet [11] and TransUNet [12] apply attention mechanism in encoder-decoder architecture to capture contextual semantics and long-range dependencies. However, simply introducing the self-attention module will increase excessive computational complexity. To solve this problem, Swin-Transformer [13] uses a hierarchical architecture to compute attention through shifting windows. Swin-UNet [14] combines the advantages of UNet and Swin-Transformer, however, it does not perform well on small-scale datasets since it does not have the inductive bias of convolution. To address this issue, UTNet [15] combines self-attention with convolution, which can fully exploit the inductive bias of convolutional networks and the long-range modeling capabilities of attention mechanisms. At the same time, the computational complexity is reduced by down-sampling key and value vectors. Since ViT [16] was proposed, transformer-based models have gradually achieved state-of-the-art (SOTA) performance for various visual tasks. However, the subsequent ConvNeXt [17] and RepLKNet [18] can also achieve comparable performance to ViT when using large depth-wise convolution kernels and following the architecture and training strategy of ViT.

In the models mentioned above, the image is regarded as a matrix with regular grid structure. However, in certain cases, expressing the image data with a graph will get better performance. Hanzhe [19] et al. proposed intra-class dynamic graph convolution to avoid interference caused by features of irrelevant categories; Yanda [20] et al. proposed to fully mine the association between region and boundary by using graph neural network (GNN) to improve the performance of semantic segmentation. However, the main issues of the existing GNNs are: (1) lack of effective relative position encoding strategies, (2) the association between nodes is temporal, and its evolution with feedforward process is not stored, (3) the information transmitted from the source node to the destination node mainly depends on the state of the source node, while the state of destination node and edge are not considered. To deal with these issues, a brain-inspired graph signal transmission (GST) network is proposed in this work. In GST, the brain neurons are represented by nodes and synapses are expressed as edges. We assume that the signal transmitted from source node to destination node must contain the following three elements: what, where, and signal strength, while satisfying the principle of signal transmission, namely the signal transmission between nodes depends on not only the their own state but also the state of the edge. Based on such assumption, we design a graph network which transmit the node and edge features, as well as the node relative positions to the next layer through a gate to control the signal strength. In the following section, we will describe in detail the structure of GST.

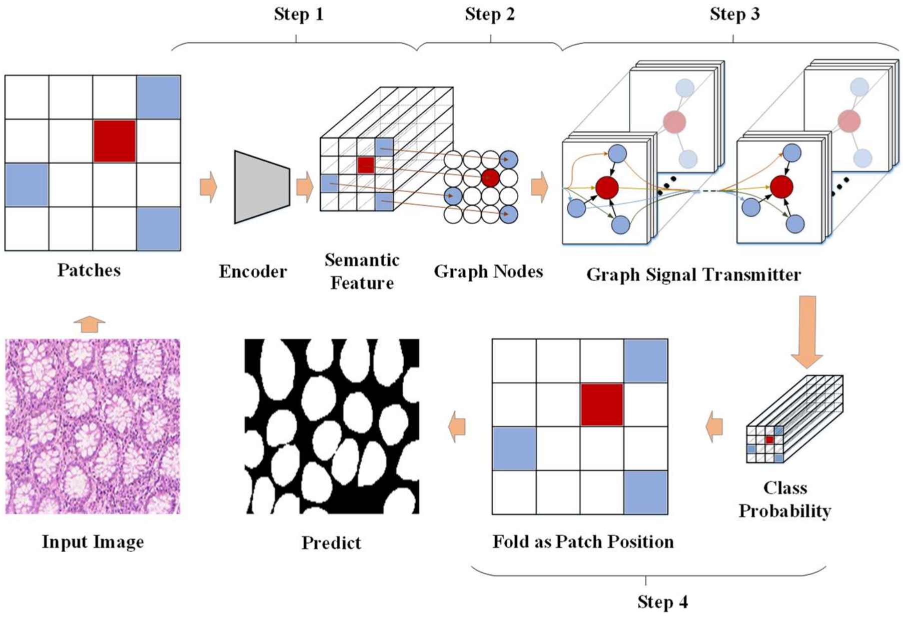


Fig. 1. The pipeline of the proposed GST for biomedical image segmentation

II. Methods

The pipeline of the proposed GST is shown in Fig.1, the image is divided into multiple patches. For simplicity, only 4×4 non-overlapping patches are shown here, the actual division method is detailed later. The GST consists of four steps. **Step1**: Extract semantic features of an image with convolution backbone encoder, and get the semantic nodes corresponding to the patches; **Step2**: Assign semantic nodes to the graph, and trim the graph according to the feature correlation between nodes; **Step3**: The signal transmission mechanism is used to communicate between the source nodes (blue) and the destination node (red), and the aggregated information from multiple source nodes is transformed by a multi-layer perceptual network (MLP) on destination node; **Step4**: The node feature vector generated in *step 3* is mapped to the class probability vector of the corresponding patch using a MLP layer, then multiple patches are spliced into the class probability map of the entire image, and finally rounded to the predicted label map. Details are described in the following subsections.

A. Construction of Semantic Nodes

The entire encoder and the bottom-level decoder layer in UNet [7] are used to extract the semantic features of graph nodes. Given the input image $\mathbf{X} \in R^{128\times192\times3}$, the resulted semantic feature is $E(\mathbf{X}) \in R^{16\times24\times512}$. Total of 16×24=384 semantic nodes are obtained, each node corresponds to a patch with size of 8 × 8 in Fig. 1, and the node feature dimension is 512. To avoid boundary effects caused by non-overlapping patches, besides the non-overlapping patch division (Fig.2(a)), three overlapping patch divisions are also used in this work (Fig.2(b)-(d)). A total of 1617 (384+425+400+408=1617) semantic nodes (patches) are obtained for each image **X**. The features of the nodes are noted as $\mathbf{H} \in R^{1617\times512}$, and the class probability of overlapping patches is averaged during prediction.

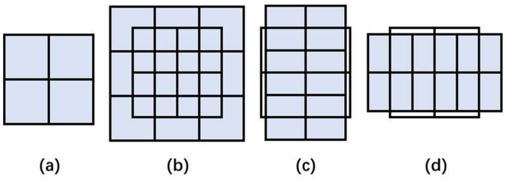


Fig. 2. Schematic diagram of overlapping patch sampling

B. Graph Structure Sampling

To reduce the computational complexity and avoid the information interference of irrelevant nodes, the association r_{ij} between destination node *i* and source node *j* is calculated based on the their node features h_i and h_j ,

$$r_{ij} = -\left\| h_i - h_j \right\| . \tag{1}$$

Only the nodes that are most relevant to the destination node (r_{ij} > threshold) are kept. As shown in Fig. 3, the red node represents the destination node, only three most relevant source nodes (blue) on the left are kept after graph sampling.

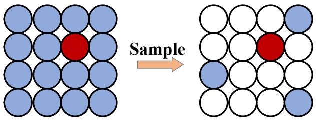


Fig. 3. Graph structure sampling

C. Signal Transmission and Transformation Mechanism

The signal transmission and transformation mechanism of GST is shown in Fig 4. The source node (Src) and the destination node (Dst) are initialized with $\mathbf{H} \in R^{1617\times512}$. Edge between node *i* and node *j* at the first layer (e_{ij}^0) is initialized with their relative position $rpos_{ij}$, defined by the linear projection of the relative coordinate $coord_{ij}$ and distance $dist_{ij}$ between nodes *i* and *j* in a cartesian coordinate system spanned by nodes.

$$rpos_{ij} = proj\left(dist_{ij} \,||\, coord_{ij}\right), \tag{2}$$

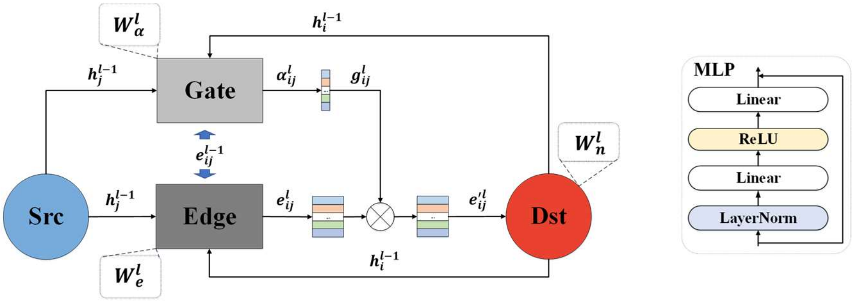


Fig. 4. Signal transmission and transformation mechanism of GST

where || indicates the concatenation operation. The source node feature h_j^{l-1}, the destination node feature h_i^{l-1}, and their edge feature e_{ij}^{l-1} of the l-1th layer are fused firstly through the concatenation and then transformed with two MLP layers to generate the transmitted signal α_{ij}^l and the edge feature e_{ij}^l of the l^{th} layer, respectively,

$$\alpha_{ij}^l = W_\alpha^l\left(h_i^{l-1} \,\|\, e_{ij}^{l-1} \,\|\, h_j^{l-1}\right), \tag{3}$$

$$e_{ij}^l = W_e^l\left(h_i^{l-1} \,\|\, e_{ij}^{l-1} \,\|\, h_j^{l-1}\right), \tag{4}$$

where W_α^l and W_e^l represent the parameters of MLP layers. This transmitted signal α_{ij}^l passes through a gate to determine how much of edge features can be transmitted to the destination node at l^{th} layer, the gate coefficient g_{ij}^l is formulated as

$$g_{ij}^l = \frac{\exp(\alpha_{ij}^l)}{\max\limits_{j\in\mathcal{N}_i}\{\exp(\alpha_{ij}^l)\}}. \tag{5}$$

Accordingly, the effective edge signal transmitted to the destination node of the l^{th} layer is written as

$$e_{ij}'^l = \tfrac{1}{N_i}\sum_{j\in\mathcal{N}_i}(g_{ij}^l e_{ij}^l) \tag{6}$$

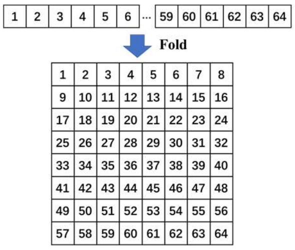
where $\mathcal{N}_i$ is the neighborhood of node i. In this work, the neighborhood is limited in the region with radius of the 8. For accurately describing the transmitted edge signal from the former layer, we divide the edge signal into K groups, the signals at each group are processed with equations (3)~(6), and the feature of destination node i at l^{th} layer is finally derived by

$$h_i^l = W_n^l\left[\mathop{\Big\|}_{k=1}^{K} e_{ijk}'^l + h_i^{l-1}\right] \tag{7}$$

where $e_{ijk}'^l$ is the effective edge signal transmitted to the l^{th} layer with the k^{th} signal group, W_n^l means the parameters of the MLP layer.

D. Segmentation in Semantic Space

We propose to use fold operation, to map directly the class probability vector of a patch regressed by MLP into the semantic space for generating a two-dimensional probability map according to the pixel positions. In Fig.5 illustrates the process of fold operation, in which, a 64-dimensional class probability vector of a given node is folded as an 8×8 patch. Folding all the nodes probability vectors results in the final class probability map.



1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

Fig. 5. Correspondence between elements of a node class probability vector and pixels of a patch.

III. Experimental Settings

A. Data Description and Preprocessings

The dataset used in this work is from GlaS Challenge Contest [1]，which has 165 H&E stained histology images and corresponding instance labels. The organizer divided it into three subsets, namely training set, test set A and test set B, which can be used to evaluate the semantic segmentation and instance segmentation performance of the model. The original image size is about 522×775, training and testing experiments are carried out with images resized to 128×192 except two comparative experiments that require the fixed input image size (Swin-UNet 224×224, UTNet 256×256). Note that, when calculating quantitative metrics for evaluating the segmentation performance of different methods, the output label map is resized to 256×384.

B. Experimental Hyperparameters and Training Strategy

We use the Pytorch framework to train and test on a single 2080Ti GPU. Optimizer is Adam, initial learning rate $lr = 0.001$, $BatchSize = 1$, $Epoch = 100$. Loss functions are MSE and IoU. The decay strategy of learning rate is *ReduceLROnPlateau* in which $patience = 3$ and $factor = 0.8$. Furthermore, data augmentation methods including random vertical and horizontal flipping, random rotation (range: −30◦ to 30◦), random deformation, random color and brightness transformation, random Gaussian blur are used.

IV. Results

A. Comparative Experimental results

In order to verify the performance of the proposed GST, it was compared with several SOTA models, including DeepLabV3 [4], UNet [7], UNet++ [8], UNet3+ [9], Swin-UNet [14], UTNet [15]. We also implemented two graph attention networks as comparative experiments on graph message passing mechanism, namely GAT [22] and graph neighbor transformer (GNT), where GNT uses Transformer's

TABLE I

THE QUANTITATIVE EVALUATION METRICS FOR GLAND SEGMENTATION WITH DIFFERENT METHODS

Models	IoU	Obj_Dice: Total/A/B			F1: Total/A/B			HD: Total/A/B			Params
GST	**0.8721**	**0.8901**	**0.9084**	**0.8353**	**0.8718**	**0.8985**	**0.7916**	**29.41**	**22.60**	**49.84**	**17.29M**
GAT	0.8498	0.8395	0.8503	0.8068	0.7764	0.8073	0.6837	45.89	38.05	69.41	17.31M
GNT	0.8524	0.8466	0.8610	0.8036	0.8003	0.8240	0.7294	42.18	35.11	63.41	17.32M
DeepLabV3	0.8091	0.7270	0.7336	0.7071	0.6434	0.6598	0.5944	82.82	75.43	104.9	15.34M
SwinUNet	0.8300	0.7529	0.7573	0.7399	0.6999	0.7171	0.6483	72.73	69.34	82.88	41.96M
UNet	0.8202	0.7881	0.7989	0.7558	0.7017	0.7355	0.6002	59.84	53.93	77.55	17.27M
UNet++	0.8235	0.8034	0.8216	0.7486	0.7233	0.7473	0.6515	58.72	49.41	86.66	47.18M
UNet3+	0.8325	0.8209	0.8376	0.7707	0.7446	0.7753	0.6526	50.01	40.91	77.31	26.98M
UTNet	0.8557	0.8288	0.8411	0.7920	0.7823	0.8038	0.7177	47.07	41.89	62.61	10.01M

TABLE II

THE QUANTITATIVE RESULTS OF DIFFERENT ABLATION EXPERIMENTS

Models	IoU	Obj_Dice: Total/A/B			F1: Total/A/B			HD: Total/A/B			Params
F(T(E(X)))	**0.8721**	**0.8901**	**0.9084**	**0.8353**	**0.8718**	**0.8985**	**0.7916**	**29.41**	**22.60**	**49.84**	**17.29M**
F(L(E(X)))	0.8485	0.8371	0.8486	0.8028	0.7836	0.8125	0.6969	45.63	38.71	66.41	13.91M
F(G(E(X)))	0.8462	0.8353	0.8465	0.8018	0.7817	0.8189	0.6701	46.16	39.90	64.93	11.27M
F(E(X))	0.8456	0.8334	0.8398	0.8144	0.7773	0.8002	0.7088	48.15	42.17	66.08	6.27M
D(E(X))	0.8362	0.7868	0.7934	0.7671	0.7258	0.7518	0.6474	60.64	55.23	76.88	6.24M
U(E(X))	0.8286	0.7844	0.7872	0.7759	0.7385	0.7519	0.6982	61.43	55.30	79.84	6.23M

encoder [23] for graph message passing. We use IoU to evaluate the semantic segmentation performance. As to the instance segmentation, three recognized metrics are used [1], including object-level Dice score (Obj_Dice), which measures the instance-level segmentation Dice coefficient, F1 score(F1), which measures the detection accuracy of gland instances, and object-level Hausdorff Distance(HD) which indicates the similarity between predicted gland boundary and that of the label. Quantitative results are shown in Table I, in which three metrics for instance segmentation are evaluated on total test set (Total), test set A (A), and test set B (B) , respectively. We find that GST outperforms the comparative models in both semantic and instance metrics. Compared with DeepLabV3, SwinUNet, UNet, UNet++, UNet3+, UTNet, GAT and GNT, GST is improved by 7.8%, 5.1%, 6.3%, 5.9%, 4.8%, 2.0%, 2.6% and 2.3% respectively on IoU; 22.4%, 18.2%, 12.9%, 10.8%, 8.4%, 7.4%, 6.0% and 5.1% on Obj_Dice; 35.5%, 24.6%, 24.2%, 20.5%, 17.1%, 11.4%, 12.3% and 8.9% on F1 score. Hausdorff Distance is reduced to 29.41, the smaller the HD is, the closer the boundary of segmentation is to the label. These results validate the superiority of GST in gland segmentation.

B. Ablation Study

To verify the effectiveness of our proposed modules, ablations of GST include: (1) Removing GST; (2) Keeping the overall architecture and replacing GST with GCN [21] and large-size depth-wise convolutions, both with the same receptive field as GST; (3) Replacing the Fold module with upsampling and transposed convolution. For the convenience of comparison, we denote the Encoder, GST, Fold, and GCN as *E, T, F, G*, respectively, and the large-size depth-wise convolution as *L*, the transposed convolution and upsampling as *D* and *U*. The quantitative results of ablation experiments are shown in Table II. The first three rows demonstrate that GST has stronger context modeling capability than large-size depth-wise convolution kernels and GCN. Compared with *F(E(X))*, *F(T(E(X)))* has an improvement of 3.1% on IoU, 6.8% on Obj_Dice, 12.2% on F1 score， and achieves a 18.7% lower HD. These results demonstrate the effectiveness of GST. The last three rows of ablation results show the superiority of the fold module over transposed convolution and upsampling.

V. DISCUSSION

As image passes through multiple convolutional layers of the encoder, the transmitted information changes from details to semantics. FCN [2] projects the feature vector in semantic space to class probability vector, and then uses transposed convolution to obtain a probability map of the same size as the original image, but the upsampling will lead to the loss of instance boundary details. As shown by the red rectangle of D(E(X)) and DeepLabV3 in Fig. 6 , the loss of details causes the two instances to stick together. To deal with this issue, the models of the UNets restore detailed information by introducing shallow features. It can be seen that the UNet family has a high degree of instance discrimination. UNet3+ achieves an instance segmentation metric of (Obj_Dice: 0.8209, F1: 0.7446, HD: 50.01), which is significantly better than the metrics (Obj_Dice: 0.7868, F1: 0.7258, HD: 60.64) of D(E(X)). Even the instance segmentation metric of UNet is generally higher than that of the model with transposed convolution.

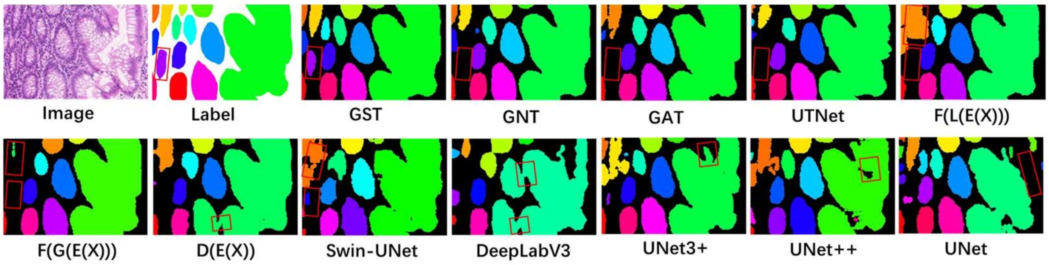


Fig. 6. Visualization of instance segmentation results of the same sample in the test set by different models.

However, we believe that shallow features can easily introduce semantically irrelevant information, which interferes with model training and prediction. For example, the IoU and instance metrics of the UNet family are lower than that of *F(E(X))* which performs segmentation directly in the semantic space. Moreover, due to the diversity of histopathological image features and the irregular shape of glands, semantically irrelevant information can easily lead to false positive in UNet3+ or false negative in UNet.

We also find that increasing the receptive field can make the boundaries more complete and smooth, such as that of GST, GAT, GNT, F(L(E(X))), F(G(E(X))) and UTNet. Conversely, the boundaries of models with small receptive field such as UNet and Swin-UNet are more fragmented. However, simply increasing the receptive field will easily cause instances with small areas to be missed, such as the red boxes shown in the lower left corner of GAT, GNT, F(L(E(X))), F(G(E(X))) and UTNet in Fig.6. Compared with large convolution kernels or self-attention, the signal transmission mechanism of GST can better model the context semantics. Furthermore, the graph structure sampling of GST makes the destination node pay attention to important source nodes with higher correlations, thus avoiding the missed detection of small instances.

VI. Conclusion

This paper proposes a new brain-inspired graph signal transmission network GST, which fully considers the principle of signal transmission: the signal transmission between nodes depends on not only the state of nodes but also the state of their edge. The relative position encoding of the source node and the destination node is introduced on the edge, and the gate is used to output the gating coefficient, so that the message passing on the Graph satisfies the three elements of what, where, and signal strength, which greatly improves the network in learning contextual semantic information. Then, the Graph is sampled based on the correlation, which not only reduces the computational complexity, but also avoids the information interference of irrelevant nodes. Finally, the node features are directly mapped back to the patch in semantic space, which not only preserves the spatial location information, but also improves the generalization of the model. Both comparative experiments and ablation experiments show that GST has significantly better segmentation performance than existing models and has potential application advantages in other medical image segmentation tasks.

Acknowledgment

This work was supported by the National Nature Science Foundations of China (Grant No.62161004), the Key Project of Guizhou Science and Technology Fund (QianKeHe ZK [2021] Key 002), Nature Science Foundations of Guizhou Province (QianKeHe [2020]1Y255).

References

[1] Sirinukunwattana, Korsuk, et al. "Gland segmentation in colon histology images: The glas challenge contest." Medical image analysis 35 (2017): 489-502.

[2] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[3] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.

[4] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

[5] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.

[6] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[7] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[8] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018. 3-11.

[9] Huang, Huimin, 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 (ICASSP). IEEE, 2020.

[10] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[11] Wang, Zhengyang, et al. "Non-local U-Nets for biomedical image segmentation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04. 2020.

[12] Chen, Jieneng, et al. "Transunet: Transformers make strong encoders for medical image segmentation." arXiv preprint arXiv:2102.04306 (2021).

[13] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[14] Cao, Hu, et al. "Swin-unet: Unet-like pure transformer for medical image segmentation." arXiv preprint arXiv:2105.05537 (2021).

[15] Gao, Yunhe, Mu Zhou, and Dimitris N. Metaxas. "UTNet: a hybrid transformer architecture for medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.

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

[17] Liu, Zhuang, et al. "A ConvNet for the 2020s." arXiv preprint arXiv:2201.03545 (2022).

[18] Ding, Xiaohan, et al. "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[19] Hu, Hanzhe, et al. "Class-wise dynamic graph convolution for semantic segmentation." European Conference on Computer Vision. Springer, Cham, 2020.

[20] Meng, Yanda, et al. "Graph-based region and boundary aggregation for biomedical image segmentation." IEEE Transactions on Medical Imaging (2021).

[21] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

[22] Veličković, Petar, et al. "Graph Attention Networks." International Conference on Learning Representations. 2018.

[23] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).