Text-Guided Cross-Position Attention for Segmentation: Case of Medical Image

Go-Eun Lee[1], Seon Ho Kim[2], Jungchan Cho[3], Sang Tae Choi[4(✉)], and Sang-Il Choi[1(✉)]

[1] Dankook University, Yongin, Gyeonggi-do, Korea
choisi@dankook.ac.kr
[2] University of Southern California, Los Angeles, USA
[3] Gachon University, Seongnam, Gyeonggi-do, Korea
[4] Chung-Ang University College of Medicine, Seoul, Korea
beconst@cau.ac.kr

Abstract. We propose a novel text-guided cross-position attention module which aims at applying a multi-modality of text and image to position attention in medical image segmentation. To match the dimension of the text feature to that of the image feature map, we multiply learnable parameters by text features and combine the multi-modal semantics via cross-attention. It allows a model to learn the dependency between various characteristics of text and image. Our proposed model demonstrates superior performance compared to other medical models using image-only data or image-text data. Furthermore, we utilize our module as a region of interest (RoI) generator to classify the inflammation of the sacroiliac joints. The RoIs obtained from the model contribute to improve the performance of classification models.

Keywords: Image Segmentation · Multi Modal Learning · Cross Position Attention · Text-Guided Attention · Medical Image

1 Introduction

Advances in deep learning have been witnessed in many research areas over the past decade. In medical field, automatic analysis of medical image data has actively been studied. In particular, segmentation which identify region of interest (RoI) in an automatic way is an essential medical imaging process. Thus, deep learning-based segmentation has been utilized in various medical domains such as brain, breast cancers, and colon polyps. Among the popular architectures, variants of U-Net have been widely adopted due to their effective encoder-decoder structure, proficient at capturing the characteristics of cells in images. Recently, it has been demonstrated that the attention modules [4,17,20] enable deep learning networks to better extract robust features, which can be applied in medical image segmentation to learn subtle medical features and achieve higher performance [14,16,18,21].

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 537–546, 2023.
https://doi.org/10.1007/978-3-031-43904-9_52

However, as image-only training trains a model with pixels that constitute an image, there is a limit in extracting fine-grained information about a target object even if transfer learning is applied through a pre-trained model. Recently, to overcome this limitation, multi-modality studies have been conducted, aiming to enhance the expressive power of both text and image features. For instance, CLIP [12] used contrastive learning based on image-text pairs to learn the similarity between the image of an object and the text describing it, achieving significant performance gains in a variety of computer vision problems.

The trend of text-image multi-modality-based research on image processing has extended to the medical field. [19] proposed a semantic matching loss that learns medical knowledge to supplement the disadvantages of CLIP that cannot capture uncertain medical semantic meaning. In [2], they trained to increase the similarity between the image and text by calculating their influence on each other as a weighted feature. For the segmentation task, LViT [10] generated the positional characteristics of lesions or target objects as text labels. Furthermore, it proposed a Double U-Shaped structure consisting of a U-Shaped ViT that combines image and text information and a U-Shaped CNN that produces a segmentation mask. However, when combining medical images with non-fine-grained text information, noise can affect the outcome.

In this paper, we propose a new text-guided cross-position attention module ($CPAM^{TG}$) that combines text and image. In a medical image, a position attention module (PAM) effectively learns subtle differences among pixels. We utilized PAM which calculates the influence among pixels of an image to capture the association between text and image. To this end, we converted the global text representation generated from the text encoder into a form, such as an image feature map, to create keys and values. The image feature map generated from an image encoder was used as a query. Learning the association between text and image enables us to learn positional information of targets in an image more effectively than existing models that learned multi-modality from medical images. $CPAM^{TG}$ showed an excellent segmentation performance in our comprehensive experiments on various medical images, such as cell, chest X-ray, and magnetic resonance image (MRI). In addition, by applying the proposed technique to the automatic RoI setting module for the deep learning-based diagnosis of sacroiliac arthritis, we confirmed that the proposed method could be effective when it is used in a practical application of computer-aided diagnosis.

Our main contributions are as follows:

- We devised a text-guided cross-position attention module ($CPAM^{TG}$) that efficiently combines text information with image feature maps.
- We demonstrated the effect of $CPAM^{TG}$ on segmentation for various types of medical images.
- For a practical computer-aided diagnosis system, we confirm the effectiveness of the proposed method in a deep learning-based sacroiliac arthritis diagnosis system.

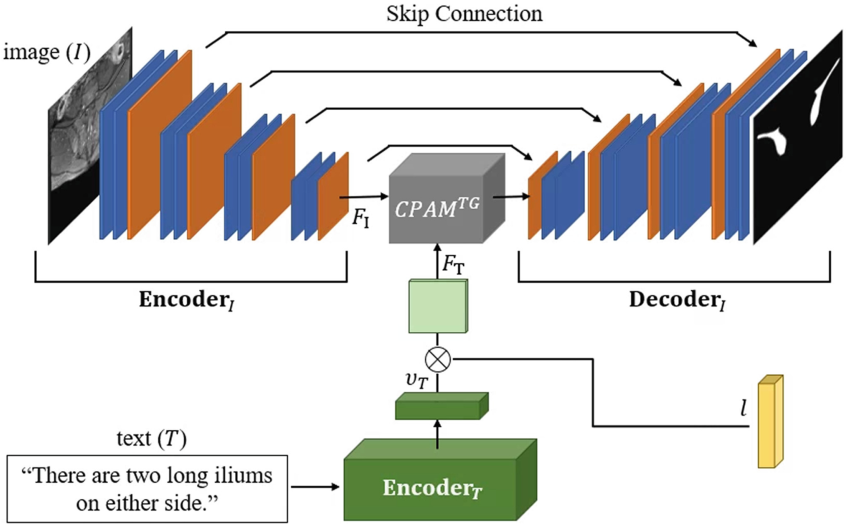

Fig. 1. Overview of our proposed segmentation model.

2 Methods

In this section, we propose text-guided segmentation model that can effectively learn the multi-modality of text and images. Figure 1 shows the overall architecture of the proposed model, which consists of an image encoder for generating a feature map from an input image, a text encoder for embedding a text describing the image, and a cross-attention module. The cross-attention module allows the text to serve as a guide for image segmentation by using the correlation between the global text representation and the image feature map. To achieve robust text encoding, we adopt a transformer [17] structure which performs well in Natural Language Processing (NLP). For image encoding and decoding, we employed U-Net, widely used as a backbone in medical image segmentation. To train our proposed model, we utilize a dataset consisting of image and text pairs.

2.1 Configuration of Text-Image Encoder and Decoder

As Transformer has demonstrated its effectiveness in handling the long-range dependency in sequential data through self-attention [1], it performs well in various fields requiring NLP or contextual information analysis of data. We used a Transformer ($\mathbf{Encoder}_T$) to encode the semantic information of the text describing a medical image into a global text representation $v_T \in \mathbb{R}^{1\times 2C}$ as $v_T = \mathbf{Encoder}_T(T)$. Here, the text semantics (T) can be a sentence indicating the location or characteristics of an interested region in an image such as a lesion shown in Fig. 1.

To create a segmentation mask from medical images (I), we used U-Net [13] which has a relatively simple yet effective structure for biomedical image segmen-

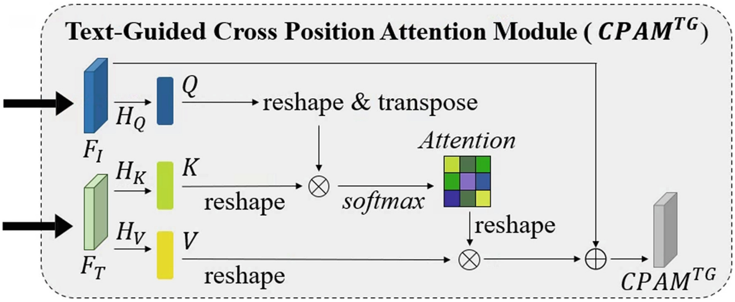

Fig. 2. Text-guided cross-position attention module ($CPAM^{TG}$).

tation. U-Net operates as an end-to-end fully connected network-based model consisting of a convolutional encoder and decoder connected by skip connections. This architecture is particularly suitable for our purpose because it can be successfully trained on a small amount of data. In the proposed method, we used VGG-16 [15] as the encoder ($\mathbf{Encoder}_I$) to obtain the image feature $F_I \in \mathbb{R}^{C\times H\times W}$ as $F_I = \mathbf{Encoder}_I(I)$ and the decoder ($\mathbf{Decoder}_I$) that will generate the segmented image from the enhanced encoding vector obtained by the cross-position attention which will be described in the following subsection.

The weights of text and image encoders were initialized by the weights of CLIP's pre-trained transformer and VGG16 pre-trained on ImageNet, respectively, and fine-tuned by a loss function for segmentation which will be described in Sect. 3.

2.2 Text-Guided Cross Position Attention Module

We introduce a text-guided cross-position attention module ($CPAM^{TG}$) that integrates cross-attention [3] with the position attention module (PAM) [5] to combine the semantic information of text and image. This module utilizes not only the image feature map from the image encoder but also the global text representation from the text encoder to learn the dependency between various characteristics of text and image. PAM models rich contextual relationships for local features generated from FCNs. It effectively captures spatial dependencies among pixels by generating keys, queries, and values from feature maps. By encoding broad contextual information into local features, and then adaptively gathering spatial contexts, PAM improves representation capability. In particular, this correlation analysis among pixels can effectively analyze medical images in which objects are relatively ambiguous compared to other types of natural images.

In Fig. 2, we multiply the learnable parameter ($l \in \mathbb{R}^{1\times(HW)}$) by the global text representation (v_T) to match the dimension of the text feature with that of the image feature map as $F_T = \mathcal{R}(G(v_T)^\intercal \times l)$, where $G(\cdot)$ is a fully connected layer that adjusts the $2C$ channel of the global text representation v_T to the image feature map channel C. $\mathcal{R}(\cdot)$ is a reshape operator to $C \times H \times W$.

The text feature map F_T is used as key and value, and the image feature map F_I is used as a query to perform self-attention as

$$Q = H_Q(F_I), \quad K = H_K(F_T), \quad V = H_V(F_T), \tag{1}$$

where H_Q, H_K, and H_V are convolution layers with a kernel size of 1, and Q, K, and V are queries, keys, and values for self-attention.

$$Attention = softmax(Q^\intercal K) \tag{2}$$

$$CPAM^{TG} = Attention^\intercal V + F_I \tag{3}$$

Finally, by upsampling the low-dimensional $CPAM^{TG}$ obtained through cross-attention of text and image together with skip-connection, more accurate segmentation prediction can express the detailed information of an object.

3 Experiments

3.1 Setup

Medical Datasets. We evaluated $CPAM^{TG}$ using three datasets: MoNuSeg [8] dataset, QaTa-COV19 [6] dataset, and sacroiliac joint (SIJ) dataset. The first two datasets are the same benchmark datasets used in [10]. MoNuSeg [8] contains 30 digital microscopic tissue images of several patients and QaTa-COV19 are COVID-19 chest X-ray images. The ratio of training, validation, and test sets was the same as in [10]. SIJ is the dataset privately prepared for this study which consists of 804 MRI slices of nineteen healthy subjects and sixty patients diagnosed with axial spondyloarthritis. Among all MRI slices, we selected the gadolinium-enhanced fat-suppressed T1-weighted oblique coronal images, excluding the first and last several slices in which the pelvic bones did not appear, and added the text annotations for the slices.

Training and Metrics. For a better training, data augmentation was used. We randomly rotated images by $-20° \sim +20°$ and conducted a horizontal flip with 0.5 probability for only the MoNuSeg and QaTa-COV19 datasets. The batch size and learning rate were set to 2 and 0.001, respectively. The loss function ($\mathcal{L}_T$) for training is the sum of the binary cross-entropy loss ($\mathcal{L}_{BCE}$) and the dice loss ($\mathcal{L}_{DICE}$): $\mathcal{L}_T = \mathcal{L}_{BCE} + \mathcal{L}_{DICE}$. The mDice and mIoU metrics, widely used to measure the performance of segmentation models, were used to evaluate the performance of object segmentation. For experiments, PyTorch (v1.7.0) were used on a computer with NVIDIA-V100 32 GB GPU.

3.2 Segmentation Performance

Table 1 presents the comparison of image segmentation performance among the proposed model and the U-Net [13], U-Net++ [22], Attention U-Net [11], MedT [16], and LViT [10] methods. Analyzing the results in Table 1, unlike natural image segmentation, the attention module-based method (Attention U-Net) and transformer-based method (MEdT) did not achieve significant performance gains compared to U-Net based methods (U-Net and U-Net++).

By contrast, LViT and $CPAM^{TG}$, which utilize both text and image information, significantly improved image segmentation performance because of multimodal complementarity, even for medical images with complex and ambiguous object boundaries. Furthermore, $CPAM^{TG}$ achieves a better performance by 1 to 3% than LViT [10] on all datasets. This means that the proposed $CPAM^{TG}$ helps to improve segmentation performance by allowing text information to serve as a guide for feature extraction for segmentation.

Figure 3 shows the examples of segmentation masks obtained using each method. In Fig. 3, we marked the boundary of the target object with a red box and showed the ground truth masks for these objects in the last column. Similar to the analysis that can be derived from Table 1, Fig. 3 shows that $CPAM^{TG}$ and LViT, which use text information together for image segmentation, create a segmentation mask with more distinctive borders than other methods. In particular, with SIJ, $CPAM^{TG}$ accurately predicted the boundaries of even thin bone parts compared to LViT. Figure 3 also shows that even on the QaTa-COV19 and MoNuSeg datasets, $CPAM^{TG}$ predicted the most accurate segmentation masks (see the red box areas). From these results, we conjecture that the reasons for the performance improvement of $CPAM^{TG}$ are as follows. $CPAM^{TG}$ independently encodes the input text and image and then combines semantic information via a cross-attention module. Consequently, the two types of information (text and image) do not act as noise from each other, and $CPAM^{TG}$ achieves an improved performance compared to LViT.

3.3 Ablation Study

To validate the design of our proposed model, we perform an ablation study on position attention and $CPAM^{TG}$. Specifically, for the SIJ dataset, we examined the effect of attention in extracting feature maps through comparison with backbone networks (U-Net) and PAM. In addition, we investigated whether text information about images serves as a guide in the position attention process for image segmentation by comparing it with $CPAM^{TG}$. Table 2 summarizes the result of each case. As can be observed in Table 2, the performance of PAM was higher than that of the backbone. This indicates that PAM improves performance by learning associations between pixels for ambiguous targets, as in medical images. In addition, the best performance results of $CPAM^{TG}$ show that text information provided helpful information in an image segmentation process using the proposed model.

Table 1. Performance comparison of medical segmentation models with three datasets

	Qata-COV19		MoNuSeg		SIJ	
	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	0.7902	0.6946	0.7645	0.6286	0.7395	0.6082
U-Net++	0.7962	0.7025	0.7701	0.6304	0.7481	0.6124
AttUNet	0.7931	0.7004	0.7667	0.6347	0.7770	0.6487
MedT	0.7747	0.6751	0.7746	0.6337	0.7914	0.6600
LViT	0.8366	0.7511	0.8036	0.6731	0.8572	0.7572
Proposed	**0.8425**	**0.7598**	**0.8105**	**0.6775**	**0.8800**	**0.7887**

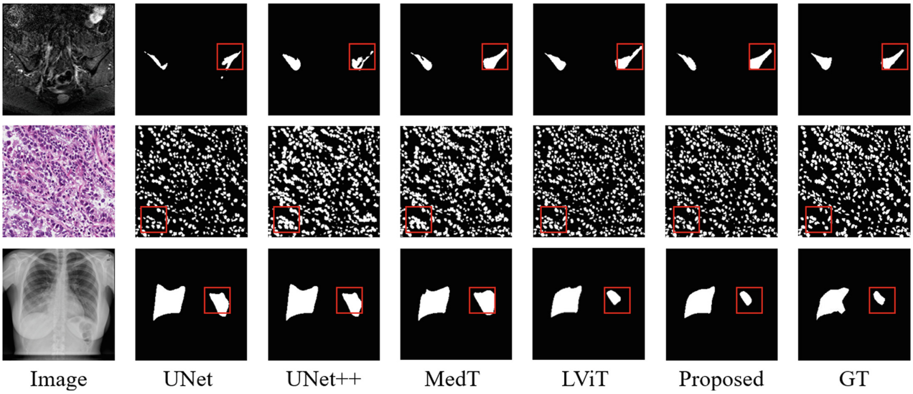

Fig. 3. Qualitative results of segmentation models.

3.4 Application: Deep-Learning Based Disease Diagnosis

In this section, we confirm the effectiveness of the proposed segmentation method through a practical bio-medical application as a deep learning-based active sacroiliitis diagnosis system.

MRI is a representative means for early diagnosis of "active sacroiliitis in axSpA". As active sacroiliitis is a disease that occurs between the pelvic bone and sacral bone, when a MR slice is input, the diagnostic system first separates the area around the pelvic bone into an RoI patch and uses it as an input for the active sacroiliitis classification network [7]. However, even in the same pelvis, the shape of the bone shown in MR slices varies depending on the slice position of the MRI and the texture of the tissue around the bone is complex. This makes finding an accurate RoI a challenge.

We segmented the pelvic bones in MRI slices using the proposed method to construct a fully automatic deep learning-based active sacroiliitis diagnosis system, including RoI settings from MRI input images. Figure 4 shows the results of generating RoI patches by dividing the pelvic bone from MRI slices using

Table 2. Ablation study of the effectiveness of our proposed module. "PAM" means we used it instead of $CPAM^{TG}$ for image-only training.

Backbone	PAM	$CPAM^{TG}$	mDice	mIoU
✓			0.7395	0.6082
✓	✓		0.7886	0.6591
✓		✓	**0.8800**	**0.7887**

Table 3. The results of classification models.

	Recall	Precision	Specificity	NPV	F1
Origin	0.8500	0.6640	0.7346	0.8880	0.7456
[9]	0.9215	0.7343	0.7424	0.9245	0.8174
Proposed	**0.9217**	**0.8281**	**0.8503**	**0.9328**	**0.8724**

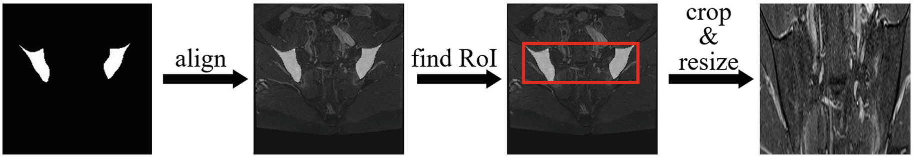

Fig. 4. Generating RoI.

the proposed method. As presented in Table 3, compared to the case of using the original MRI image without the RoI setting, using the hand-crafted RoI patch [9] showed an average of 7% higher performance in recall, precision, and f1. It is noticeable that the automatically set RoI patch showed similar or better performance than the manual RoI patch for each measurement. This indicates that the proposed method can be effectively utilized in practical applications of computer-aided diagnosis.

4 Conclusion

In this study, we developed a new text-guided cross-attention module ($CPAM^{TG}$) that learns text and image information together. The proposed model has a composite structure of position attention and cross-attention in that the key and value are from text data, and the query is created from the image. We use a learnable parameter to convert text features into a tensor of the same dimension as the image feature map to combine text and image information effectively. By calculating the association between the reshaped global text representation and each component of the image feature map, the proposed method outperformed image segmentation performance compared to previous studies using both text and image or image-only training method. We also confirmed that it could be utilized for a deep-learning-based sacroiliac arthritis diagnosis system, one of the use cases for practical medical applications. The proposed method can be further used in various medical applications.

Acknowledgements. This work was supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program (RS-2022-00155227) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2021R1A2B5B01001412), and the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (RS-2023-00220408).

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Bhalodia, R., et al.: Improving pneumonia localization via cross-attention on medical images and reports. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 571–581. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_53
3. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366 (2021)
4. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
6. Haghanifar, A., Majdabadi, M.M., Choi, Y., Deivalakshmi, S., Ko, S.: Covid-cxnet: detecting covid-19 in frontal chest x-ray images using deep learning. Multimedia Tools Appl. **81**(21), 30615–30645 (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging **36**(7), 1550–1560 (2017)
9. Lee, K.H., Choi, S.T., Lee, G.Y., Ha, Y.J., Choi, S.I.: Method for diagnosing the bone marrow edema of sacroiliac joint in patients with axial spondyloarthritis using magnetic resonance image analysis based on deep learning. Diagnostics **11**(7), 1156 (2021)
10. Li, Z., et al.: Lvit: language meets vision transformer in medical image segmentation. arXiv preprint arXiv:2206.14718 (2022)
11. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)

13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Shome, D., et al.: Covid-transformer: interpretable covid-19 detection using vision transformer for healthcare. Int. J. Environ. Res. Public Health **18**(21), 11086 (2021)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
17. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)
18. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
20. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
21. Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: nested modality-aware transformer for brain tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 140–150. Springer, Heidelberg (2022)
22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1