COARSE-TO-FINE COVID-19 SEGMENTATION VIA VISION-LANGUAGE ALIGNMENT

Dandan Shan[⋆] *Zihan Li*[†] *Wentao Chen*[‡] *Qingde Li*[∗] *Jie Tian*[◇] *Qingqi Hong*[⋆]

[⋆]Xiamen University [†]UIUC [‡]BUPT [∗]University of Hull [◇]CAS

ABSTRACT

Segmentation of COVID-19 lesions can assist physicians in better diagnosis and treatment of COVID-19. However, there are few relevant studies due to the lack of detailed information and high-quality annotation in the COVID-19 dataset. To solve the above problem, we propose C2FVL, a Coarse-to-Fine segmentation framework via Vision-Language alignment to merge text information containing the number of lesions and specific locations of image information. The introduction of text information allows the network to achieve better prediction results on challenging datasets. We conduct extensive experiments on two COVID-19 datasets including chest X-ray and CT, and the results demonstrate that our proposed method outperforms other state-of-the-art segmentation methods.

***Index Terms*—** Coarse-to-Fine, Vision-Language, Semantic Segmentation

1. INTRODUCTION

Deep learning technique has been widely used in the area of medical image processing, and the segmentation of lesions using neural networks can significantly reduce time and labor costs. Since multimodal images exhibit more relevant information [1] than unimodal images, some studies [2, 3] merge images of different modalities to segment the target. However, medical images lack detailed information, making it challenging to segment precisely only by medical images. Some works [4, 5, 6] use the method of image information fusion with text information for natural image segmentation. Although Tomar et al. [7] applied image and text information fusion to medical image segmentation, it did not consider the location information of target. The segmentation of the lesion area is an effective means to diagnose and treat COVID-19. However, only experienced radiologists can accurately label lesions, and with few publicly available datasets [8], make it difficult to improve the accuracy of segmentation networks [9, 10]. To solve the above problems, we introduce text information in the process of COVID-19 segmentation. Since the text contains the number and specific location information of lesions, it can assist the network in learning more features with rich semantic information from coarse to fine on a limited and poorly labeled dataset. In summary, the main contributions of this paper are:

- We construct the C2FVL segmentation framework using CNN and Vision Transformer to achieve accurate segmentation of COVID-19 effectively.
- We propose a Vision Language Alignment Module (VLAB) and a novel loss function to facilitate the alignment of text and image information, which improves segmentation accuracy.
- We compare our C2FVL with the state-of-the-art segmentation methods on two COVID-19 datasets, and the experiments show that the performance of C2FVL is optimal. The code of our proposed C2FVL is made available at GitHub[1].

2. RELATED WORKS

Multi-modal Information Fusion: Multiple modality can provide complementary information, and their fusion can effectively improve the accuracy of the segmentation of medical images. Dolz et al. [2] used dense connections between different modalities to swap information. CLIP [4] used a contrast learning approach to learn image and text information, achieving zero-shot transfer learning. TGANet [7] introduced text about the size and number of polyps, enabling automatic segmentation of polyps of different scales.

COVID-19 Segmentation: Saeedizadeh et al. [9] added a regularization term to the loss function of the UNet to detect chest regions infected with COVID-19 in CT images. Alom et al. [10] used the NABLA-N model to segment the COVID-19-infected region in the designed system. Yao et al. [11] alleviate the problem of difficult-to-acquire annotation by learning contextual knowledge from normal lungs but there is still a performance gap compared to the supervised approach.

3. METHOD

3.1. Structure of C2FVL Model

As shown in Fig. 1, our model follows an encoder-decoder structure, where the encoder uses hierarchical CNN and Vision Transformer to extract coarse image features. At the skip

The first two authors have the equal contribution.

Corresponding author: Qingqi Hong.

[1]https://github.com/HUANGLIZI/C2FVL

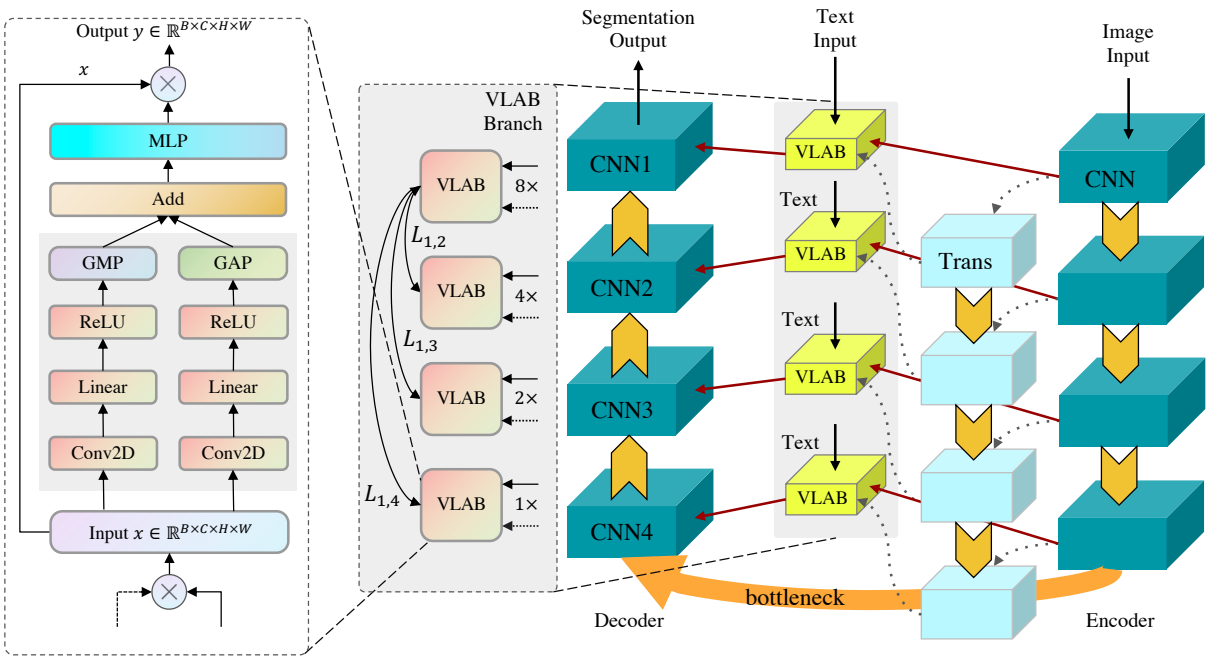


Fig. 1. Overview of the C2FVL framework. It consists of three parts, encoder, VLAB branch and decoder.

connection, the text information is fused with image information and design the VLAB module to facilitate cross-modal alignment between text and images. The decoder consists of CNN that processes low-level detail information and aligned text-image features to reconstruct fine segmentation masks.

3.2. Encoder Module

Traditional convolutional neural networks use a fixed-size receptive field to learn to local features of an image. In contrast, the Multi-headed Self-attention mechanism in the Vision Transformer works directly on the whole image and can learn more global features. For medical image segmentation tasks, both global and local features are essential. Therefore, we use several CNN and Vision Transformer blocks in the encoder module to extract image features jointly. First, we take the image as the input of the CNN block and then send the feature map to the Vision Transformer after convolution operation, batch normalization, and activation function. At the same time, max-pooling is used for downsampling the feature map, which is sent to the next layer of CNN block. For the activation function, we use ReLU to enhance the expression ability of the network. In the Vision Transformer part, the features extracted by the CNN are encoded to learn global features further. At each stage, the scale of the output of Vision Transformer is adjusted to the same as the output of the CNN block, and adding them together to fuse local and global features, as shown in the following equations.

$$y_{rt,i} = \sigma\left(BatchNorm\left(Conv\left(Upsample\left(y_{vit,i}\right)\right)\right)\right) \quad (1)$$

$$y_{encoder,i} = y_{cnn,i} + y_{rt,i} \quad (2)$$

where $y_{vit,i}$ denotes the output of the i-th layer Vision Transformer, $y_{rt,i}$ denotes the result of the reconstruction of $y_{vit,i}$, and $y_{cnn,i}$ denotes the output of the i-th layer CNN, σ is ReLU activation function.

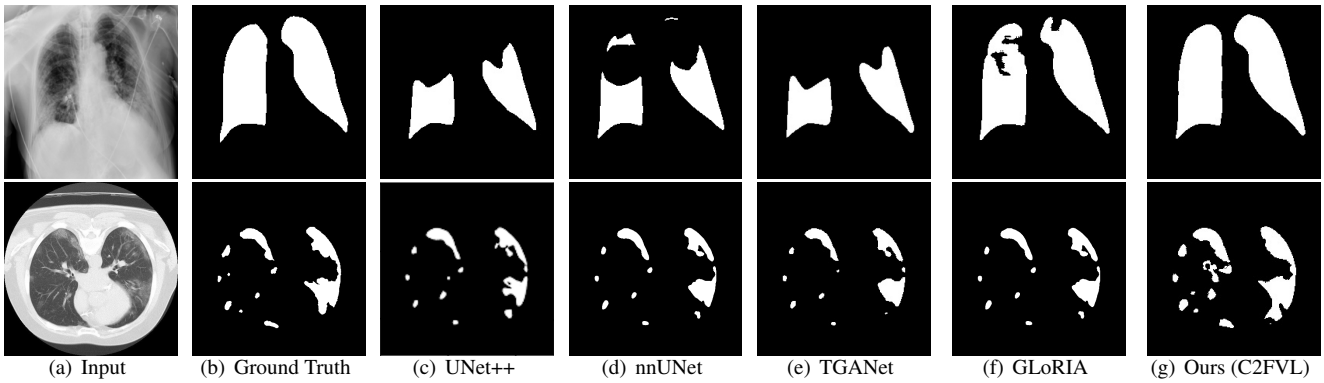
3.3. Multi-sclale Vision-Language Aggregation

We convert the text information into a vector with dimension of 8. The first dimension indicates whether the lesion area is bilateral, the second dimension indicates the number of lesions, and each subsequent three dimensions indicates the location of the lesions in the left and right lungs in that order. For example, suppose the text message is "Bilateral pulmonary infection, two infected areas, upper middle left lung and middle lower right lung". In that case, we transform it into a vector $V_{text} = [1, 2, 1, 1, 0, 0, 1, 1]$. Unlike [7], for feature maps of different scales, we employ the repeat operation to align V_{text} with the number of channels of the image features and then weight the image features with text to extract the lesion's location and quantity features, while suppressing other irrelevant features, as shown in Eqn. (3).

$$F_{vl,i} = y_{encoder,i} \odot \left(V_{text}.repeat\left(channel_i/8, 1\right)\right) \quad (3)$$

3.4. Vision-Language Alignment Block (VLAB)

Inspired by the convolutional block attention module [12], we design our Vision-Language Alignment Block (VLAB) with a parallel structure to facilitate the alignment of text and image features. As shown in Fig. 1, in the left and right branches, the input features are first processed using the MLP structure and then the feature maps are subjected to global max pooling (GMP) and global average pooling (GAP), respectively. The outputs of two branches are added to further enhance the same



(a) Input (b) Ground Truth (c) UNet++ (d) nnUNet (e) TGANet (f) GLoRIA (g) Ours (C2FVL)

Fig. 2. Visualization results on the QaTa-COVID dataset (row 1) and MosMedData+ dataset (row 2). From left to right: (a) input image, (b) ground-truth, (c) UNet++ and (d) nnUNet are predictions of baseline without text information, while (e) TGANet and (f) GLoRIA are predictions with text information. And (g) is the prediction of our proposed C2FVL.

features, and finally multiplied with the input after two layers of MLP structure. The process can be defined as follows.

$$F_{avg} = GlobalAvgPool\left(MLP\left(x\right)\right) \tag{4}$$

$$F_{max} = GlobalMaxPool\left(MLP\left(x\right)\right) \tag{5}$$

$$y = MLP\left(F_{avg} + F_{max}\right) \otimes x \tag{6}$$

In addition, We also calculate the cosine loss between different VLAB outputs to force each network layer to focus on the same focal region and append this part of the loss after the dice loss L_{Dice} and the cross entropy loss L_{CE}, with the designed loss function defined as shown in Eqn. (7) and (8).

$$L_{4,i} = 1 - \frac{y_4 \cdot Downsample(y_i)}{|y_4| \times |Downsample(y_i)|} \tag{7}$$

$$L_{V1} = \frac{1}{2}L_{Dice} + \frac{1}{2}L_{CE} + \alpha L_{4,1} + \beta L_{4,2} + \gamma L_{4,3} \tag{8}$$

where $L_{4,i}$ $(i \in 1, 2, 3)$ denotes the cosine loss between the 4-th layer VLAB output y_4 and the i-th layer VLAB output y_i. The default values of α, β, γ are all 0.5. In addition, we also design $L_{1,i}$ $(i \in 2, 3, 4)$ to calculate the cosine loss between the outputs of the first layer VLAB and the output of the lower layer VLAB.

$$L_{1,i} = 1 - \frac{y_1 \cdot Upsample(y_i)}{|y_1| \times |Upsample(y_i)|} \tag{9}$$

$$L_{V2} = \frac{1}{2}L_{Dice} + \frac{1}{2}L_{CE} + \alpha L_{1,4} + \beta L_{1,3} + \gamma L_{1,2} \tag{10}$$

4. EXPERIMENTS

4.1. Setup

Datasets: We use open-acess QaTa-COVID dataset [13] and MosMedData+ dataset [14, 15] in the experiments to evaluate the performance. The training and test sets of QaTa-COVID contain 7145 and 2113 chest X-ray images with annotations, respectively. We use a 4:1 ratio to split the initial training set into training and validation. MosMedData+ contains 2729 CT scan slices of lung infections. To balance the different morphologies of the lesion, we consider the images with four prefixes as four classes. For each class, we distribute them in a ratio of 8:1:1 in the training set, validation set, and test set.

Implementation Details: Our experiments are conducted on four NVIDIA GeForce GTX 1080 Ti GPUs with the initial learning rate of model training set to 1e-3 and the batch size to 4. The total number of iterations is set to 2000 rounds, and the number of early stop rounds is set to 50.

Table 1. Performance comparison between our method (C2FVL) and other state-of-the-art methods on datasets.

Method	Text	QaTa-COVID		MosMedData+	
		Dice (%)	IoU (%)	Dice (%)	IoU (%)
UNet [16]	×	79.02	69.46	64.60	50.73
UNet++ [17]	×	79.62	70.25	71.75	58.39
AttUNet [18]	×	79.31	70.04	66.34	52.82
nnUNet [19]	×	79.89	70.58	72.59	60.36
TransUNet [20]	×	78.63	69.13	71.24	58.44
Swin-UNet [21]	×	77.27	67.96	63.29	50.19
UCTransNet [22]	×	79.15	69.60	65.90	52.69
TGANet [7]	✓	79.87	70.75	71.81	59.28
GLoRIA[23]	✓	79.94	70.68	72.42	60.18
Ours	✓	**83.40**	**74.62**	**74.56**	**61.15**

4.2. Comparison with SOTA Methods

We conducted comparison experiments on the two datasets with state-of-the-art methods, including methods with only image input and methods with image-text input. As shown in Fig. 2, the first row shows that other methods occur in the upper part of the lung with different missing sizes, but our proposed C2FVL model can segment the lesion relatively completely. As seen in the second row, the results predicted by C2FVL are closer to the ground truth for the right lung region, which is more important for diagnosis. As shown in Table 1, comparing the best-performing method nnUnet without text, C2FVL has a 3.51% higher Dice score and 4.04% higher IoU on the Qata-COVID dataset. Meanwhile, on the MosMedData+ dataset, the Dice score and IoU are improved by 1.97% and 0.79%, respectively. Compared with the baseline model with text like GLoRIA [23], C2FVL achieves all improvements of 3.46%, 3.94%, 2.14% and 0.97% under these two datasets in both two metrics.

4.3. Ablation Study

We performe a series of ablation experiments on the QaTa-COVID dataset for the associated hyper-parameters and different components of the model.

Ablation study on different hyper-parameters: For batch size, we selected three values of 2, 4 and 8 for comparison, and for learning rate, we set 1e-4, 1e-3, 1e-2. From the results in Table 2, we can see that when batch size is 4, learning rate is 1e-3, the model has the best Dice score and IoU value. For the loss coefficient, we start with $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.5$, fixing two values among α , β, γ as 0.5, and change the other value to observe trend of the model performance. As shown in Fig. 3, the optimal set of parameters is $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.5$.

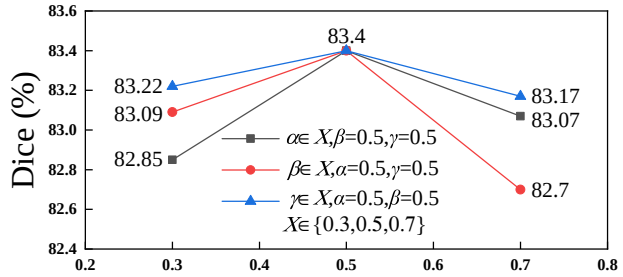


Fig. 3. Ablation study with different cosine loss coefficients. The gray, red and blue lines represent the trend of dice with α, β and γ as variables, respectively.

Table 2. Ablation study on different hyper-parameters.

Hyper-Parameters	Value	QaTa-COVID	
		Dice (%)	IoU (%)
Batch Size	2	82.03	73.32
	4	**83.40**	**74.62**
	8	83.32	74.57
Learning Rate	1e-4	82.93	74.19
	1e-3	**83.40**	**74.62**
	1e-2	82.83	73.92

Table 3. Ablation study on effectiveness of different components on QaTa-COVID dataset.

Method	CNN	Text Input		C2FVLoss		VLAB	Dice (%)	IoU (%)
		Single	Multi	L_{V1}	L_{V2}			
UNet++	✓						79.62	70.25
C2FVL	✓	✓					83.01	74.26
	✓		✓				83.18	74.43
	✓		✓	✓		✓	83.08	74.42
	✓		✓		✓	✓	**83.40**	**74.62**

Ablation study on effectiveness of different components: As shown in Table 3, we use CNN that extracts image features as the backbone and then gradually added text information and VLAB. It is worth noting that when the single text is added to the original image input, the model's performance was substantially improved compared to the backbone, with Dice score improving by 3.39% and IoU by 4.01%. The model's performance is further improved after adding the corresponding scale text in each layer. On this basis, the performance of VLAB with cosine loss using the downsampling method decreases, but the Dice score and IoU values of VLAB with cosine loss using upsampling method reach the optimum. Therefore, we use L_{V2} for the VLAB in the subsequent experiments.

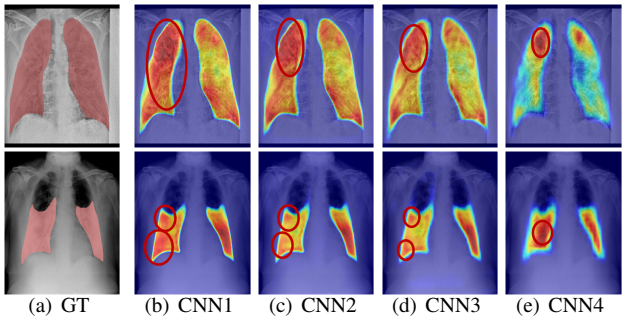


Fig. 4. Saliency map for interpretability study on QaTa-COVID dataset. From (a) to (d) are the saliency maps corresponding to the outputs of four CNN decoders from bottom (CNN4) to top (CNN1) of different outputs in each CNN decoder. (e) "GT" represents the corresponding ground truth.

4.4. Interpretability Study

As shown in Fig. 4, to verify that our cosine loss enhances the network's focus on the lesion region, we use Grad-CAM [24] to show the activation region of the decoder part on the QaTa-COVID dataset. As seen from the left lung in the first row, the activation mapping of CNN4 shows a relatively narrow region of strong activation. As the layers of the network become progressively shallower, the region of strong activation begins to spread from the location of CNN4 to nearby regions, eventually focusing on the entire lesion region. For the left lung in the second row, the activation map of CNN4 shows the strong activation region mainly at the center of the lesion region, then splits to the edge of the left lung, followed by spreading from the edge to the central region. This result also demonstrates that the designed cosine loss in C2FVL can help the model detect the location of inaccurate focus and force it to refocus on the correct lesion region.

5. CONCLUSION

In this paper, we propose a new segmentation network called C2FVL, which uses text and image information in the training process. In addition, we design a VLAB to facilitate the alignment of image and text information and add cosine loss between different layers in the loss function to make each layer of the network focus on the lesion region. The experimental results show that our model improves performance compared to other state-of-the-art models. While our current work scratches the surface on multi-scale Vision Transformers for image classification, we anticipate that in future there will be more works in developing C2FVL for more applications, including more new released datasets.

6. REFERENCES

[1] Binny Naik et al., "Denouements of machine learning and multimodal diagnostic classification of alzheimer's disease," *Visual Computing for Industry, Biomedicine, and Art*, vol. 3, no. 1, pp. 1–18, 2020.

[2] Jose Dolz, Karthik Gopinath, et al., "Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.

[3] Xiaohang Fu, Lei Bi, et al., "Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3507–3516, 2021.

[4] Alec Radford, Jong Wook Kim, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[5] Jiasen Lu, Dhruv Batra, et al., "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[6] Weijie Su, Xizhou Zhu, Yue Cao, et al., "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[7] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, et al., "Tganet: Text-guided attention for improved polyp segmentation," *arXiv preprint arXiv:2205.04280*, 2022.

[8] Sneha Kugunavar et al., "Convolutional neural networks for the diagnosis and prognosis of the coronavirus disease pandemic," *Visual computing for industry, biomedicine, and art*, vol. 4, no. 1, pp. 1–14, 2021.

[9] Narges Saeedizadeh et al., "Covid tv-unet: Segmenting covid-19 chest ct images using connectivity imposed unet," *Computer Methods and Programs in Biomedicine Update*, vol. 1, pp. 100007, 2021.

[10] Md Zahangir Alom, MM Rahman, Mst Shamima Nasrin, Tarek M Taha, and Vijayan K Asari, "Covid_mtnet: Covid-19 detection with multi-task deep learning approaches," *arXiv preprint arXiv:2004.03747*, 2020.

[11] Qingsong Yao, Li Xiao, Peihang Liu, and S Kevin Zhou, "Label-free segmentation of covid-19 lesions in lung ct," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2808–2819, 2021.

[12] Sanghyun Woo, Jongchan Park, et al., "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[13] Mehmet Yamac et al., "Convolutional sparse support estimator-based covid-19 recognition from x-ray images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 1810–1820, 2021.

[14] Sergey P Morozov, AE Andreychenko, et al., "Mosmeddata: Chest ct scans with covid-19 related findings dataset," *arXiv preprint arXiv:2005.06465*, 2020.

[15] "Covid-19 ct segmentation dataset," [EB/OL], `http://medicalsegmentation.com/covid19/` Accessed December 23, 2020.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

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

[18] Ozan Oktay, Jo Schlemper, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[19] Fabian Isensee, Paul F Jaeger, et al., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[20] Jieneng Chen, Yongyi Lu, Qihang Yu, et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[21] Hu Cao, Yueyue Wang, Joy Chen, et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[22] Haonan Wang, Peng Cao, et al., "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2441–2449.

[23] Shih-Cheng Huang, Liyue Shen, et al., "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3942–3951.

[24] Ramprasaath R Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.