MTNET: MULTI-SCALE TEXT-AWARE NETWORK FOR COVID-19 SEGMENTATION

Zihan Li[⋆†] *Dandan Shan*[⋆] *Yuan Zheng*[⋆] *Qingqi Hong*[⋆] *Qingqiang Wu*[⋆]

[⋆] Xiamen University [†]UIUC

ABSTRACT

Most of the existing COVID-19 segmentation methods usually apply other multi-organ segmentation methods to COVID-19 segmentation. However, these works ignore the distribution and size of COVID-19 lesions are relatively uncertain, compared with organs. It is difficult to carry out high-precision segmentation only depending on image features, which is exactly what we want to pursue. Therefore, we propose a new multi-scale text-aware network (MTNet) to address the above issues. We utilize multi-scale text information to help locate the lesions, and image-text joint learning to further improve the segmentation performance. The evaluation of two COVID-19 datasets shows that our method can achieve state-of-the-art performance compared with other methods. The code will be available on GitHub.

Index Terms— COVID-19 Segmentation, Text-aware, Multi-scale Learning

1. INTRODUCTION

In recent years, COVID-19 has become a global pandemic and has affected the lives and bodies of many people. And it is always a challenging problem to segment COVID-19 lesion accurately due to its uncertainty that means lesions may exist **anywhere** in the lung in **any size**. However, current study usually does not consider the above issues together. Many researcher [1, 2, 3, 4] mainly focus on how to separate the infected area from the healthy area but fail to segment small lesions. Although [3] has paid attention to the lesion boundary, segmentation performance is still limited by the image quality and lack of annotation. As for multi-scale learning, although many works[5, 6, 7] employ convolutional kernels of different sizes to handle multi-scale features, it is hard to segment lesion regions at different scales due to the irregular morphology and hard-to-detect noise. Specifically, for small size targets, background is often wrongly classified as foreground, while for large size targets, foreground is usually regarded as background. According to our solution, to tackle the problem of "anywhere", we turn our attention to the medical reports. We hold the view that clinical information from the medical reports could be help to localize the position of lesion, so we decide to explore the information to get a better segmentation performance. To address the issue of "any size", we plan to apply text feature to image feature of corresponding scales, and uniformly consider information of different scales.

Our main contributions are summarized as follows:

- We propose a new COVID-19 segmentation framework (MTNet), to the best of our knowledge, which is the first framework to take multi-modal and multi-scale problems into consideration jointly.
- We propose the Multi-scale Text-aware Block (MT-Block) and Image-Text Joint Learning to perceive image and text information simultaneously in the network.
- We evaluate the performance of MTNet on two datasets (COVID-Xray and COVID-CT) and results show that our MTNet achieves superior performance over the existing state-of-the-art segmentation methods.

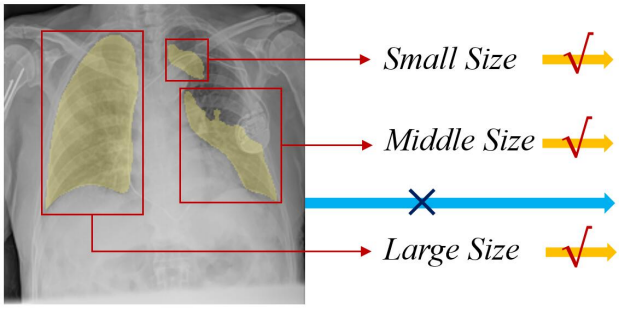

Fig. 1. Distribution of COVID-19. The lesions may be distributed in different locations with different sizes, our MTNet considers different location and different scale jointly.

2. RELATED WORKS

COVID-19 Segmentation: Deep learning technology can assist in COVID-19 screening and automatically segment infected regions. Fan et al.[1] proposed a semi-supervised segmentation framework to segment COVID-19 infected regions from CT slices. The existing semi-supervised segmentation methods [2] ignore the geometric constraints, which leads to the non-smooth boundary of object. To resolve this issue, Huang et al.[3] used a semi-supervised segmentation framework with boundary detection capability. However, traditional deep learning methods often require a large amount of computation and are difficult to be deployed in practice.

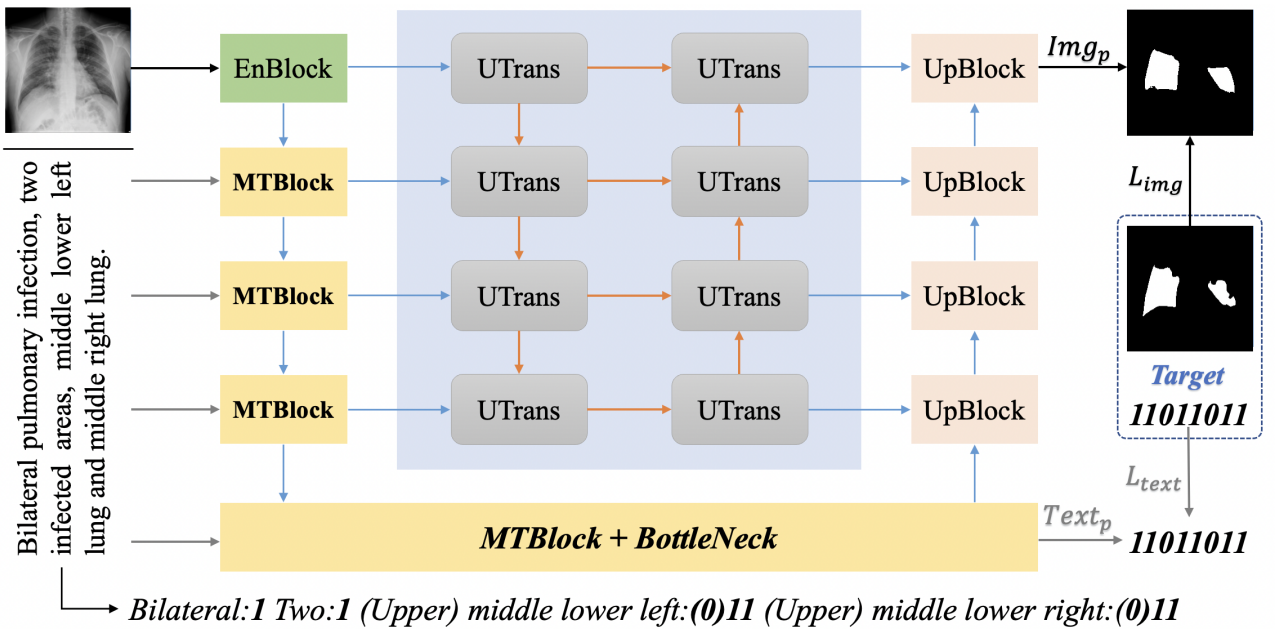

Fig. 2. Overview of our proposed MTNet. The framework consists of the hybrid CNN-Transformer architecture. Specifically, MTBlock is used to receive text and image features of the same scale, and transmits features to UTrans for secondary encoding. Finally, the image segmentation prediction and text one-hot prediction are jointly supervised by the mixed loss.

Multi-scale Learning: Features at different scales contain different scale information, and the features of different scales is essential for segmentation. Yan et al.[5] proposed Progressive Atrous Spatial Pyramid Pooling (PASPP), which uses combination of Atrous convolutions. MSD-Net[6] uses pyramid convolution blocks to deal with multi-scale problem of lesion segmentation. Pei et al.[7] designed a multiscale feature extraction structure and sieve connection module.

3. METHOD

3.1. Overview of MTNet

An overview of our framework MTNet is shown in Fig.2. Here, we will briefly introduce the whole framework. The MTNet consists of several CNN blocks such as MTBlock, EnBlock and UpBlock and Transformer blocks such as UTrans. Each block can represent different features and transmit the features to each other. First, the image is encoded as the image features through convolution, batch normalization operations inside EnBlock. Next, MTBlock receives both image features from upper layer and corresponding scale text features. It is worth noting that the text features are converted from text by using one-hot encoding as shown in Fig.2. Then, we construct a u-shaped structure (UTrans) to further improve the representation ability of mixed features including image features and text features. Owing to the belief of image features and text features of same scales will have similarity to a certain extent in representation, we unify them to the same scale and utilize transformer structure for secondary encoding to enhance the similarity. Finally, the image segmentation prediction from UpBlock and text one-hot prediction from BottleNeck are jointly supervised by mixed loss L_{img}+L_{text}. And different from the previous multimodal methods[8, 9, 10], MTNet does not have a multimodal decoder for cross modality fusion, but performs multimodal fusion in the same encoder. We believe the integration of text information will help image encoding in this way.

3.2. Multi-scale Text-aware Block (MTBlock)

We design a new multi-scale Text-aware block (MTBlock) to do interaction between image feature $F_{img} \in \mathbb{R}^{C,H,W}$ and text feature $F_{text} \in \mathbb{R}^{H,W}$. And F_{img} is encoded by EnBlock first as shown in Eqn. 1. As for F_{text}, we encode the text with one-hot encoding, e.g. the description "Bilateral pulmonary infection, two infected areas, middle lower left lung and middle right lung." is encoded as '11011011'. Then the binary code $Text_B$ '11011011' is converted to the decimal code $Text_D$ '219'. In order to align the scale of different feature, we construct the matrix $\mathbf{1}$ of corresponding size, and then dot product it with $Text_D$ to obtain F_{text}. In addition, we set the scale factor η to regulate F_{text} as shown in Eqn. 2. Then F_{img} is concated with F_{text} to form F_{mix}. As shown in Eqn. 4, we perform the MP (MaxPooling), BN (Batch Normalization) and Convolution operation on F_{mix} to form the input $F^{'}_{mix}$ of next MTBlock.

$$F_{img} = EnBlock(Img) \tag{1}$$

$$F_{text} = \frac{\mathbf{1} \odot Text_D}{\eta} \tag{2}$$

$$F_{mix} = [F_{img}; F_{text}] \tag{3}$$

$$F^{'}_{mix} = Conv(BN(MP(F_{mix}))) \tag{4}$$

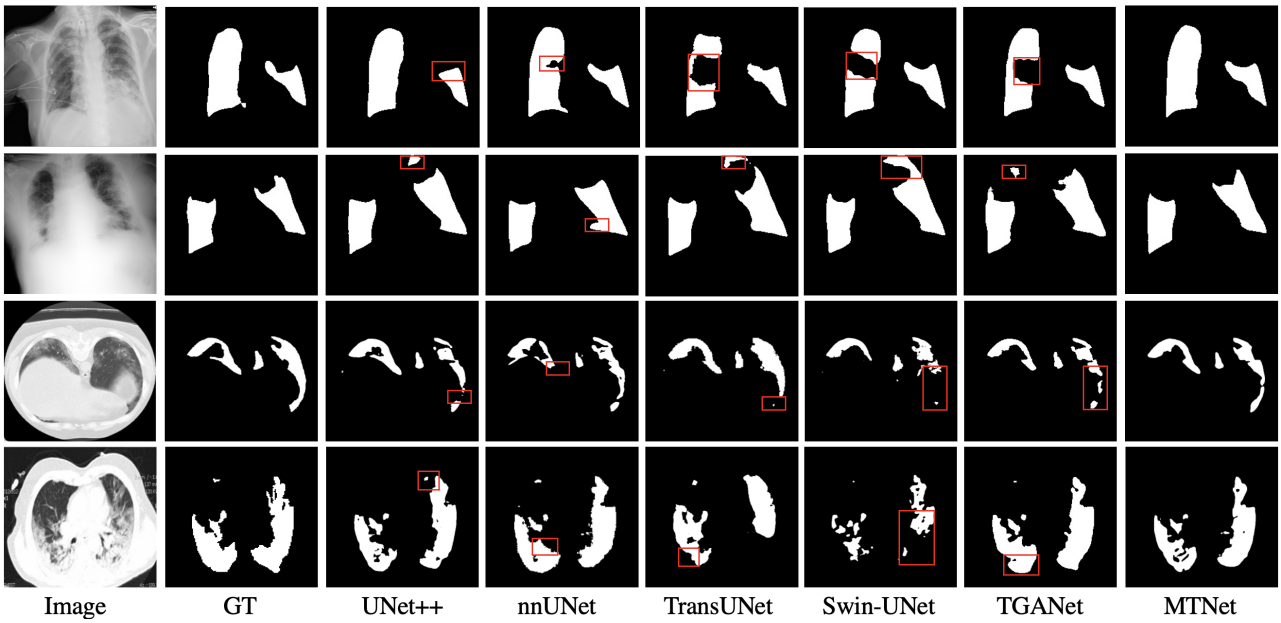

Fig. 3. Qualitative results on COVID-Xray and COVID-CT datasets. Red boxes indicate differences between GT and prediction.

3.3. Image-Text Joint Learning

As shown in Fig.2, image-text joint learning is introduced into our pipeline. Different loss functions are utilized on image output and text output for supervision. The image loss L_{img} consists of the Dice Loss and Cross Entropy Loss as shown in Eqn. 5, where p_i represents the prediction result of the i-th pixel, and y_i represents the ground truth (GT) value of the corresponding pixel. And we also apply the Cross Entropy Loss as the text loss L_{text} to supervise text classification, where p_t represents the text prediction $Text_p$ of the t-th sample, and y_t represents the GT value of the same sample. Besides, we add the coefficient α of L_{text} as shown in Eqn. 7.

$$\mathrm{L}_{img} = 1 - \sum_{i=1}^{N} \left(\frac{1}{N} \cdot \frac{2|p_i \cap y_i|}{(|p_i| + |y_i|)} + \frac{1}{N} \cdot y_i \log\left(p_i\right)\right) \quad (5)$$

$$\mathrm{L}_{text} = -y_t \log\left(p_t\right) \quad (6)$$

$$\mathrm{L}_{mix} = \mathrm{L}_{img} + \alpha \times \mathrm{L}_{text} \quad (7)$$

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Datasets and Implementation Details

COVID-Xray. [11] The dataset contains a total of 9258 chest X-ray images, the training set contains 7145, the test set contains 2113, and radiologists label the Covid-19 infection area of each image. **COVID-CT.** [12, 13] This dataset combines 2729 slices from three publicly available datasets, each of which was labeled by different radiologists with different labels, and the different datasets were combined to uniformly map different colored masks to white.

Implementation Details. For the COVID-Xray dataset, we adopt the official test set and divide the 20% of data in the official train set as validation set. For the COVID-CT dataset, we randomly divide the train set, the validation set and the test set according to the ratio of 8:1:1. We implement our method based on the PyTorch framework and 2-card RTX A5000. We set the initial learning rate as 3e-4 and batch size as 24. The number of early stop epoch we set is 100.

4.2. Comparison with Other State-of-the-Art Methods

In Table 1, we compare our proposed MTNet method with the other eleven state-of-the-art methods. Our method is superior to other methods in Dice score and IoU metric on COVID-Xray and COVID-CT datasets. Especially on the COVID-CT dataset, the Dice coefficient and IoU metric of our method are increased by 2.19% and 1.51% respectively compared with nnUNet[14]. As shown in Fig. 3, the first and second rows indicate the segmentation results for the COVID-Xray dataset, and the third and fourth rows show the segmentation results for the COVID-CT dataset. The qualitative comparison results show our proposed MTNet's superiority over other state-of-the-art methods. Specifically, the first and third rows of the figure show that the segmentation of the left and right lungs by other networks occurs with varying degrees of missingness, while MTNet produces segmentation results close to ground truth, with relatively complete segmentation for the problematic segments. From the segmentation results shown in the second and fourth rows of the figure, the other baseline networks show fragments of different sizes. However, our proposed MTNet achieves a relatively good performance. Overall, the fusion of multi-scale features effectively reduces incorrect segmentation and decreases the occurrence of false positive and false negative rates.

Table 1. Performance comparison between our proposed method (MTNet) and other state-of-the-art methods.

Method	Text	COVID-Xray		COVID-CT	
		Dice(%)	IoU(%)	Dice(%)	IoU(%)
UNet[15]	×	79.02	69.46	64.60	50.73
UNet++[16]	×	79.62	70.25	71.75	58.39
AttUNet[17]	×	79.31	70.04	66.34	52.82
nnUNet[14]	×	79.90	70.59	72.59	60.36
TransUNet[18]	×	78.63	69.13	71.24	58.44
Swin-UNet[19]	×	77.27	67.96	63.29	50.19
UCTransNet[20]	×	79.15	69.60	65.90	52.69
ConVIRT[8]	✓	79.72	70.58	72.06	59.73
GLoRIA[10]	✓	79.94	70.68	72.42	60.18
TGANet[21]	✓	79.87	70.75	71.81	59.28
MTNet (ours)	✓	**80.97**	**71.97**	**74.78**	**61.87**

4.3. Ablation Studies

In this subsection, we conduct several experiments on the COVID-Xray dataset to validate the effectiveness of our proposed model, including the hyperparameter settings and each key component.

Ablation studies of different Hyper-Parameters. We evaluate the effect of different scale factors of Text input and different coefficients of L_{text} on the segmentation performance. For the scale factor, we compared the results for four values of 32, 64, 128, 256. For the coefficient of Ltext, we set three values of 0.01,0.05,0.1. According to Table. 2, scale factor of 32 and coefficient of 0.05 are optimal for the Dice coefficient and IoU values.

Ablation studies of different components. Ablation experiments were conducted on the critical components of the proposed method, and the relevant experimental results are shown in Table. 3. We gradually add MTBlock, Text, UTrans and L_{text} to the baseline. From the experimental results, all four improvements are effective in performance improvement, and adding MTBlock improves the average Dice score by 0.96% compared to the baseline. On this basis, by adding text information and UTrans in turn, the accuracy of the model has been further improved. Finally, with the introduction of L_{text}, the performance is significantly improved compared to the baseline. Specifically, Dice and IoU reached 80.97% and 71.97%, respectively.

Table 2. Ablation studies of MTNet on different Hyper-Parameters: Scale Factor of Text input, Coefficient of L_{text}.

Hyper-Parameters		**Xray**: Dice(%)	**Xray**: IoU(%)
Scale Factor: η	32	**80.97**	**71.97**
	64	80.55	71.59
	128	80.51	71.61
	256	80.37	71.40
Coefficient: α	0.01	80.29	71.25
	0.05	**80.97**	**71.97**
	0.10	80.36	71.39

Table 3. Ablation studies of different components: MTBlock, Text, UTrans, L_{text} on the COVID-Xray dataset. The second row indicates that only MTBlock is evaluated, and additional text input is replaced by the original image input.

MTBlock	Text	UTrans	L_{text}	Dice(%)	IoU(%)
◯	◯	◯	◯	79.02	69.46
✓	◯	◯	◯	79.78	70.33
✓	✓	◯	◯	80.35	71.32
✓	✓	✓	◯	80.59	71.47
✓	✓	✓	✓	**80.97**	**71.97**

4.4. Multi-scale Text-aware Analysis

To verify the effectiveness of Multi-scale text input, we observe the effect of text input of different layers on the segmentation performance. Fig.4 shows four different operations for different layers of text input. As the blue line shows, our proposed method of incorporating the corresponding text features in each layer of image features can effectively improve performance with optimal results. As can be observed from the orange and grey lines, the improvement in performance of the model is more pronounced by incorporating text input at a shallow layer, which can be further explored in the future.

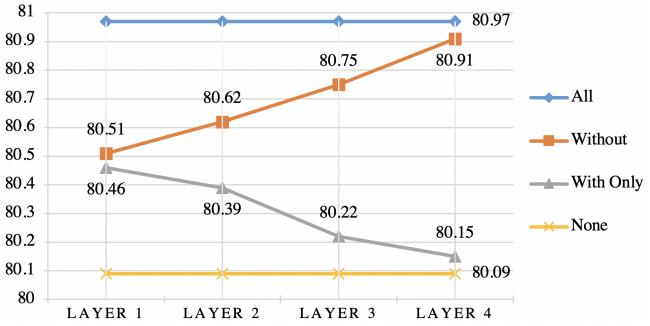

Fig. 4. Analysis of different layers of text input. 'All' represents our proposed MTNet, 'With Only' represents only the text input of corresponding layer is kept and 'Without' represents only the text input of corresponding layer is removed. 'None' represents without any text input.

5. CONCLUSION

In this paper, MTNet is proposed to improve the performance of COVID-19 segmentation. The MTBlock is utilized to receive feature and interact different modality in MTNet. In addition, we propose the image-text joint learning to achieve the consistency of supervision at the image and text level. Experimental results show that MTNet achieves comparable segmentation performance on the two COVID-19 datasets. We also explore the differences in text input at different layers. We find the performance improvement of model is more pronounced by incorporating the text input at a shallow layer.

6. REFERENCES

[1] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, et al., "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[2] Shixuan Zhao, Zhidan Li, Yang Chen, et al., "Scoat-net: a novel network for segmenting covid-19 lung opacification from ct images," *Pattern Recognition*, vol. 119, pp. 108109, 2021.

[3] Huimin Huang, Lanfen Lin, Yue Zhang, et al., "Graphbas3net: Boundary-aware semi-supervised segmentation network with bilateral graph convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7386–7395.

[4] Zihan Li, Dihan Li, Cangbai Xu, et al., "Tfcns: A cnn-transformer hybrid network for medical image segmentation," in *International Conference on Artificial Neural Networks*. Springer, 2022, pp. 781–792.

[5] Qingsen Yan, Bo Wang, Dong Gong, et al., "Covid-19 chest ct image segmentation network by multi-scale fusion and enhancement operations," *IEEE transactions on big data*, vol. 7, no. 1, pp. 13–24, 2021.

[6] Bingbing Zheng, Yaoqi Liu, Yu Zhu, Fuli Yu, Tianjiao Jiang, Dawei Yang, and Tao Xu, "Msd-net: Multi-scale discriminative network for covid-19 lung infection segmentation on ct," *Ieee Access*, vol. 8, pp. 185786–185795, 2020.

[7] Hong-Yang Pei, Dan Yang, Guo-Ru Liu, and Tian Lu, "Mps-net: Multi-point supervised network for ct image segmentation of covid-19," *Ieee Access*, vol. 9, pp. 47144–47153, 2021.

[8] Yuhao Zhang, Hang Jiang, et al., "Contrastive learning of medical visual representations from paired images and text," *arXiv preprint arXiv:2010.00747*, 2020.

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

[10] 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.

[11] Aysen Degerli, Mete Ahishali, Mehmet Yamac, et al., "Covid-19 infection map generation and detection from chest x-ray images," *Health information science and systems*, vol. 9, no. 1, pp. 1–16, 2021.

[12] Jun Ma, Yixin Wang, Xingle An, et al., "Towards efficient covid-19 ct annotation: A benchmark for lung and infection segmentation," 2020.

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

[14] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, 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.

[15] 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.

[16] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, 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.

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

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

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

[20] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane, "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.

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