MAdapter: A Better Interaction between Image and Language for Medical Image Segmentation

Xu Zhang[1], Bo Ni[2], Yang Yang[3] (✉), and Lefei Zhang[1] (✉)

[1] School of Computer Science, Wuhan University, Wuhan, China
zhanglefei@whu.edu.cn
[2] Computer School, Hubei Polytechnic University, Huangshi, China
[3] Renmin Hospital, Wuhan University, Wuhan, China
carlayangyang@whu.edu.cn

Abstract. Conventional medical image segmentation methods are only based on images, implying a requirement for adequate high-quality labeled images. Text-guided segmentation methods have been widely regarded as a solution to break the performance bottleneck. In this study, we introduce a bidirectional **M**edical **A**daptor (MAdapter) where visual and linguistic features extracted from pre-trained dual encoders undergo interactive fusion. Additionally, a specialized decoder is designed to further align the fusion representation and global textual representation. Besides, we extend the endoscopic polyp datasets with clinical-oriented text annotations, following the guidance of medical professionals. Extensive experiments conducted on both the extended endoscopic polyp dataset and additional lung infection datasets demonstrate the superiority of our method. The code and text annotation are available at https://github.com/XShadow22/MAdapter.

Keywords: Medical Image Segmentation · Multi-Modal Learning · Vision-Language

1 Introduction

Computer-aided diagnosis (CAD) has become integral to medical studies, with semantic segmentation serving as a fundamental process. Existing segmentation methods, whether based on Unet [1] or transformer [2] architectures, are designed for single-modality, thus imposing greater demands on sufficient high-quality labeled images. With the development of Natural Language Processing(NLP), researchers have recognized the potential of using language as supplementary supervision signals to tackle the challenge of insufficient annotated images. The clinical textual reports accompanying medical images often indicate the segmentation objects as well as their quantity, size, position, and other relevant information. This helps compensate for deficiencies in image quality and enhance segmentation performance.

Previous representative study Clip [10] constructed a multi-modal model based on contrastive pre-training on sufficient image-text pairs. [6,7,8] extended the Clip model to downstream segmentation tasks. Some studies [9] have also validated the effectiveness of using textual information to guide medical image segmentation. Lvit [3] introduced a hybrid CNN-transformer structure to integrate textual and visual information. Lee et al. [12] utilized a text-guided cross-position attention mechanism. However, their cross-modality fusion strategy is relatively simple. Zhong et al. [4] employed GuideDecoder blocks to enable multi-modal information fusion.

However, in previous medical image segmentation methods, textual semantic information unidirectionally propagated into visual features, only acting as supervisors during the fusion process. When there is representation bias in itself, it will introduce textual semantic noise, leading to a decrease in segmentation performance. Thus we consider facilitating multi-scale mutual information interaction to overcome this limitation. In this paper, we introduce a bidirectional **A**daptor for **M**edical image segmentation (MAdapter) to connect multi-layers of uni-modal encoders. By integrating independent MAdapters to facilitate interaction between encoders, we inject task-specific information into the fixed backbone while preserving its inherent feature extraction capability. We also propose a specialized decoder to globally align information and ultimately generate dense predictions. To validate the adaptability of our proposed method, we extend the endoscopic polyp datasets with text annotations and conduct experiments on the extended polyp datasets and two other datasets, totally consisting of three different image modalities(endoscopic images, X-ray, and CT). Our contributions can be summarized as follows:

- We propose a cross-modal medical image segmentation framework, using additional text annotations to enhance segmentation performance.
- We introduce MAdapter and a specialized decoder, facilitating bidirectional interactive fusion and alignment of multi-level visual and linguistic features.
- We extend the endoscopic polyp datasets with detailed and clinical-oriented text annotations.
- Extensive experiments are conducted on the extended datasets and other publicly medical image-text benchmark datasets. All results demonstrate the superiority of our proposed method.

2 Method

In this section, we provide a detailed description of the proposed cross-modal medical image segmentation method. We first utilize fixed vision and language encoders to extract feature maps. A bidirectional MAdapter is proposed to make the interaction of visual and linguistic information. Finally, dense predictions are generated through a lightweight decoder where combined features and global textual representations are aligned. Adjusting only the parameters of the MAdapter and decoder allows for flexible adaptability and injection of task-specific knowledge. The architecture of the proposed method is illustrated in Figure 1.

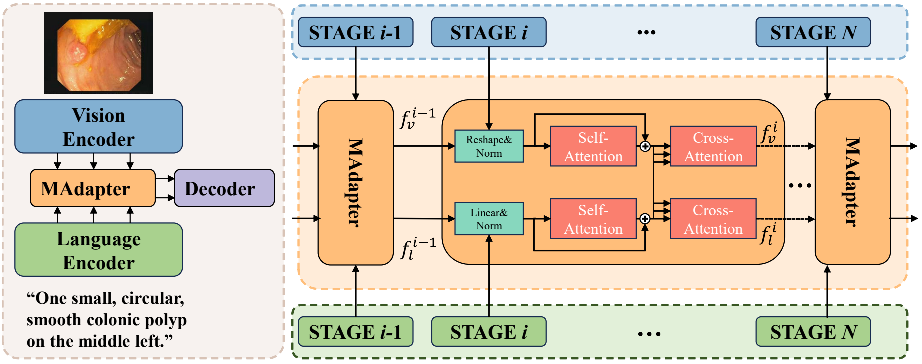

Fig. 1. The overview of our proposed method (left) and the detailed MAdapter structure (right). Built upon a fixed pre-trained vision and language model, our approach incorporates trainable interaction modules and a decoder, which can be easily integrated into existing models. The MAdapter facilitates information interactive fusion.

2.1 Image and Text Feature Extraction

For an input image $I \in \mathcal{R}^{H \times W \times 3}$, we extract visual features from different stages of the image encoder, denoted as $F_v^i, i \in \{1, ..., N\}$, where N represents the number of stages, which corresponds to the number of MAdapters. For an input text prompt $T \in R^L$, similar to the image encoder, we extract linguistic features and a global textual feature from the language encoder, represented as $F_l^i \in \mathcal{R}^{L \times C}, i \in \{1, ..., N\}$ and f_g. The multi-level visual and language features from different blocks will be employed in our framework as the input of the MAdapter and decoder for multi-modal feature interaction.

2.2 Image and Text Feature Interaction

MAdapter. The vision and language encoders are mutually invisible. However, thanks to the carefully designed MAdapter module, our proposed framework can propagate additional knowledge to individual modality information through bidirectional interaction.

Given multiple visual features $F_v^i, i \in \{1, ..., N\}$ and linguistic features $F_l^i, i \in \{1, ..., N\}$, we obtain enhanced information through MAdapter. We firstly reshape the visual features and project the linguistic features onto an appropriate dimensionality using a linear layer, followed by passing them through a layer normalization step. This process can be formalized as

$$\begin{aligned} \tilde{F}_v^i &= Norm(Reshape(f_v^{i-1})) + F_v^i \\ \tilde{F}_l^i &= Norm(Linear(f_l^{i-1})) + F_l^i \end{aligned} \tag{1}$$

where f_v^{i-1}, f_l^{i-1} represent the feature map obtained from the previous MAdapter. The interaction module we employ is based on an attention mechanism. Initially, we apply self-attention modules separately to process visual and textual

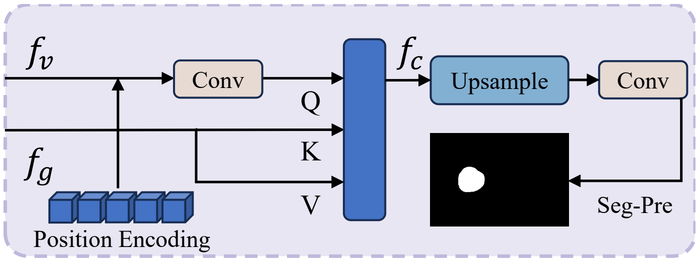

Fig. 2. The overview of the lightweight decoder. It globally aligns textual information and fine-grained combined features, ultimately generating segmentation predictions.

information, incorporating a residual connection to obtain enhanced features. Subsequently, we utilize a cross-attention module. Specifically, we set the Query (Q) to one modality's feature and designate the Keys (K) and Values (V) to the other modality's feature. The process can be formalized as

$$\begin{aligned} \tilde{F}_v^i, \tilde{F}_l^i &= \mathcal{F}_{MHCA}(\tilde{F}_v^i) + \tilde{F}_v^i, \mathcal{F}_{MHCA}(\tilde{F}_l^i) + \tilde{F}_l^i \\ f_v^i, f_l^i &= \mathcal{F}_{MHCA}(\tilde{F}_v^i, \tilde{F}_l^i), \mathcal{F}_{MHCA}(\tilde{F}_l^i, \tilde{F}_v^i) \end{aligned} \tag{2}$$

where $\mathcal{F}_{MHCA}$ denotes multi-head cross-attention. After the fusion of N layers of MAdapter, the final combined features containing local textual semantic information are denoted as f_v.

Decoder. To predict masks for each pixel, we employ a lightweight decoder. The previous MAdapter performed multi-scale local information fusion and output cross-modal fine-grained feature f_v. In the decoder, the global textual representation f_g is introduced to make alignment with f_v. We start by incorporating positional encoding into mixed features. Then, utilizing cross-attention modules, we derive multi-modal features f_c to capture comprehensive global contextual information. After upsampling, convolutional layers, and activation functions σ, f_c is transformed into the final segmentation results Pre. The process can be formalized as:

$$\begin{aligned} \tilde{f}_v &= Conv(f_v) \\ f_c &= \mathcal{F}_{MHCA}(\tilde{f}_v, f_g) \\ \tilde{f}_c &= Upsample(f_c) \\ Pre &= \sigma Conv(\tilde{f}_c) \end{aligned} \tag{3}$$

3 Experiments

3.1 Datasets

We evaluate our method on two segmentation tasks. One task is the segmentation of lung infection areas. The QaTa-COV19 dataset [13] contains 9258 chest X-ray images of COVID-19 cases and the MosMedData+ [14] dataset contains 2729 CT scan slices of lung infections. Thanks to Li et al. [3] for their work in providing text annotations.

The other task is the classical polyp segmentation task. We selected five public polyp segmentation datasets, including CVC-ClinicDB [16], Kvasir [17], ETIS [18], ColonDB [19], and CVC-300 [20]. Following [15], the images from Kvasir and ClinicDB are randomly selected for training. The detailed dataset information is listed in the supplementary material.

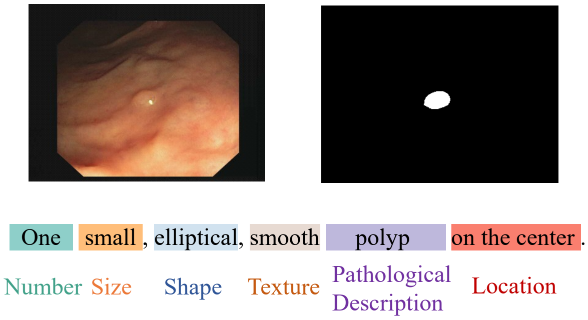

Fig. 3. A typical annotated case with different descriptors highlighted in distinct colors.

It is worth noting that the five original datasets did not come with corresponding textual annotations. Therefore, we extended them under the guidance of medical professionals. Specifically, the annotation has been organized into six categories, including the pathological description of the lesion name, the number of lesion areas present, the size of each lesion area, the texture of the lesions, the shape of the lesions, and the location of each lesion area. Compared to [3], our text annotations encompass a more wide range of pathological descriptions. This helps provide more comprehensive semantic information. Figure 3 illustrates a typical annotated case.

3.2 Experiment Settings

Our model is trained on a workstation with two NVIDIA RTX 3090 GPUs using the PyTorch framework. We use the AdamW as our optimizer. The initial learning rate lr_0 is set to 5e-3. We choose a batch size of 48 for the lung infection segmentation and 24 for the polyp segmentation. All images are uniformly preprocessed, including standard intensity normalization and cropping to 224×224. The pre-trained vision and language models we used are ConvNeXt-Tiny [22] and CXR-BERT [23] respectively. We use the dice and cross-entropy loss as the loss function. Two metrics are used to evaluate the segmentation results: the Dice and mIoU coefficients, which are widely used for segmentation tasks. The detailed description is listed in the supplementary material.

3.3 Results

We compare the segmentation performance of our model with UNet [1], Unet++ [5], BiomedClip [21], Lvit [3], and LGMS [4]. It is worth noting that the first two methods are single-modal segmentation models based on images only. The latter three methods are multi-modal models based on Clip, CNN-transformer hybrid architecture, and fine-tuning of pre-trained models, respectively. The quantitative results for lung infection region segmentation are presented in Table 1, while those for the polyp segmentation are shown in Table 2.

Table 1. Performance comparison of different models for lung infection segmentation.

Method	QaTa-COV19		MosMedData+	
	Dice	mIoU	Dice	mIoU
Unet	79.31	69.93	65.81	51.75
UNet++	79.81	71.11	71.35	58.22
BiomedClip	87.88	78.38	66.51	50.52
Lvit	83.66	75.11	74.57	61.33
LGMS	89.87	81.15	77.34	63.06
Ours	**90.22**	**82.16**	**78.62**	**64.78**

With the introduction of textual information, the segmentation performance of the model has been significantly improved. Compared to existing multi-model methods, we demonstrate superior performance in lung infection region segmentation tasks. This improvement is attributed to the carefully designed interaction module, enhancing the transferability of powerful pre-trained vision and language models.

By extending the corresponding text annotations of the endoscopic images dataset, the proposed model could be evaluated on the classic polyp segmentation task. The result presented in Table 2 reveals that, compared to some methods with complex structures and cumbersome parameters, our model achieves comparable or even superior performance with an easy and versatile structural design.

Table 2. Performance comparison of different models for polyp segmentation.

Method	ClinicDB		Kvasir		ColonDB		CVC-300		ETIS	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
UNet	82.49	75.61	81.82	74.61	52.33	45.58	71.35	62.84	39.82	33.52
Unet++	79.49	72.95	82.51	74.36	48.35	41.07	70.74	62.45	40.15	34.47
BiomedClip	83.61	76.51	85.66	77.91	67.40	50.83	72.89	63.25	52.46	40.89
Lvit	89.20	82.54	90.13	82.51	73.12	57.62	84.53	71.88	63.24	53.80
LGMS	93.87	88.45	91.16	83.76	77.82	63.70	90.31	82.34	82.36	70.01
Ours	**95.46**	**91.32**	**91.58**	**84.47**	**78.85**	**65.80**	**91.94**	**85.09**	**82.85**	**70.72**

The qualitative experimental results are illustrated in Figure 4. It can be observed that our model achieves more accurate edge detection compared to both single-modal and multi-modal methods.

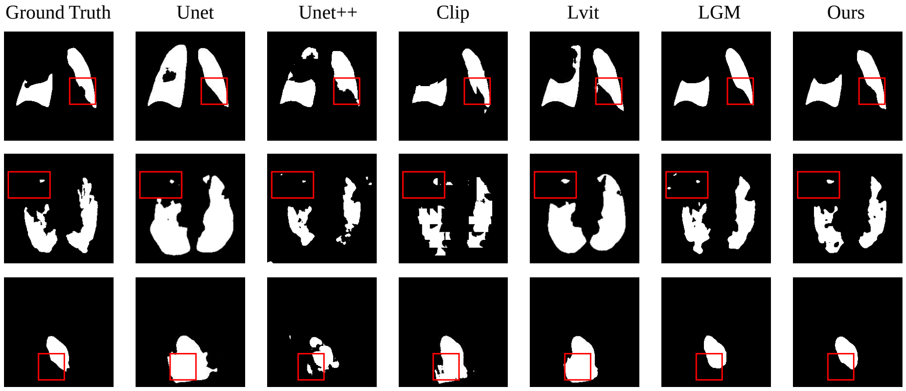

Fig. 4. Qualitative results on the QaTa-COV19, the MosMedData+ and ClinicDB datasets. The red boxes indicate that we have achieved better edge detection and segmentation results.

3.4 Ablation Study

Effectiveness of Proposed Components: Our proposed method introduced textual information and enhanced cross-modal information interaction and correction. We conducted ablation studies to demonstrate the effectiveness of the text and each component. Specific details are provided in Table 3.

Table 3. Ablation studies on QaTa-COV19, MosmedData+, ClinicDB test set. 'w/o text' means without text and the model uses vision decoder only. 'MAdapter' means the model utilizes MAdapter for modality interaction but employs only a simple segmentation head. 'MAdapter+DC' is our proposed method.

Method	QaTa-COV19		MosMedData+		ClinicDB	
	Dice	mIoU	Dice	mIoU	Dice	mIoU
w/o text	82.82	70.68	74.30	60.01	89.48	80.96
MAdapter	89.74	81.38	78.15	64.13	95.01	90.49
MAdapter+DC	**90.22**	**82.16**	**78.62**	**64.78**	**95.46**	**91.32**

From Table 3, taking results on QaTa-Cov19 as an example, it can be observed that the Dice metric improved by 6.92% with the introduction of textual information, demonstrating the effectiveness of interaction facilitated by

MAdapter. Additionally, with the upgrade of the decoder, the Dice metric increased by 0.48%, indicating the benefits of global alignment.

Ablation Study on Different Sets of Prompts: A medical report will provide multi-faceted descriptions of pathology. To investigate the impact of different text granularities on segmentation results, we dissected the text prompts. P1 represents 'number, name', P2 represents 'number, size, name', P3 represents 'number, size, shape, name', P4 represents 'number, size, shape, texture, name', and full-text represents 'number, size, shape, texture, name, location'. Table 4 presents the evaluation results on ClinicDB and CVC300.

Table 4. Study of different sets of text prompts.

Datasets	Metrics	P1	P2	P3	P4	Full-Text
ClinicDB	Dice	93.85	94.62	95.25	94.71	**95.46**
	mIoU	88.42	89.78	90.93	89.95	**91.32**
CVC300	Dice	90.89	**91.98**	91.22	91.52	91.94
	mIoU	83.31	**85.15**	83.85	84.36	85.09

Through our experiments, we observe that richer descriptions lead to improvements in segmentation performance. With the full-text prompts configuration, the Dice metric increased by 1.61% and 1.05%, respectively, compared to the simplest P1 setting. However, we also note that different medical descriptors do not contribute equally to the results. Descriptions related to quantity and localization have a more significant impact. This may be due to their inherent objectivity and clear evaluation criteria. Conversely, descriptors such as texture rely more on subjective judgment, which could potentially lead to the introduction of wrong information and consequently result in negative effects.

4 Conclusion

In this paper, we introduce a novel multi-modal framework for medical image segmentation. Specifically, we utilize a bidirectional MAdapter to facilitate interaction between multi-level visual and textual information extracted from fixed dual-encoders. A flexible decoder is employed to refine the alignment between the global textual representation and the combined representation. These individually tuned modules can be seamlessly integrated into any pre-trained model, injecting task-specific information while retaining its inherent feature extraction capability. Furthermore, we extend the corresponding textual annotations for the endoscopic polyp dataset. The proposed method is validated on the extended polyp dataset and several other benchmark datasets, demonstrating its superiority and adaptability. demonstrating its superiority and adaptability.

Acknowledgments. This project is funded by the Cross-Innovative Talent Project of RenMin Hospital of Wuhan University (Project No: JCRCZN-2022-006).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28
2. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)
3. Li, Z. et al., LViT: Language Meets Vision Transformer in Medical Image Segmentation. IEEE Transactions on Medical Imaging, **43**(1), 96-107, (2024)
4. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's Thread: Using Text Prompts to Improve Segmentation of Infected Areas from Chest X-ray Images. In: Greenspan, H., et al. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, vol 14223. Springer, Cham. https://doi.org/10.1007/978-3-031-43901-8_69
5. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging, **39**(6), 1856-1867 (2019).
6. Lüddecke, T. and Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7086-7096 (2022).
7. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11686-11695 (2022).
8. Xu, Z., Chen, Z., Zhang, Y., Song, Y., Wan, X., Li, G.: Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17503-17512 (2023).
9. Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., Khanal, B.: Exploring transfer learning in medical image segmentation using vision-language models. arXiv preprint arXiv:2308.07706(2023)
10. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
11. Liu, C. et al.: M-FLAG: Medical Vision-Language Pre-training with Frozen Language Models and Latent Space Geometry Optimization. In: Greenspan, H., et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, vol 14220. Springer, Cham. https://doi.org/10.1007/978-3-031-43907-0_61
12. Lee, GE., Kim, S.H., Cho, J., Choi, S.T., Choi, SI. : Text-Guided Cross-Position Attention for Segmentation: Case of Medical Image. In: Greenspan, H., et al.(eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, vol 14224. Springer, Cham. https://doi.org/10.1007/978-3-031-43904-9_52

13. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: OSegNet: operational segmentation network for Covid-19 detection using chest X-ray images. In: Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), pp. 2306–2310. IEEE (2022)
14. Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzymyrskyy, A.V., Ledikhova, N.V., Gombolevskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.V., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465(2022)
15. Fan, DP. et al.: PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In: Martel, A.L., et al.(eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science(), vol 12266. Springer, Cham. https://doi.org/10.1007/978-3-030-59725-2_26
16. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. CMIG **43**, 99–111 (2015)
17. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
18. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014)
19. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging **35**(2), 630–644 (2015)
20. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M. and Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering, **2017** (2017)
21. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M.P.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915.(2023)
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
23. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision-language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XXXVI, pp. 1–21. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1