Towards Robust Medical Image Referring Segmentation with Incomplete Textual Prompts

Qijie Wang[1], Xian Lin[1], and Zengqiang Yan[1(✉)]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology
{wqijie, xianlin, z_yan}@hust.edu.cn

Abstract. Recent advancements in medical vision-language models have increasingly accentuated the substantial potential of incorporating textual information for better medical image segmentation. However, existing language-guided segmentation models were developed under the assumption that the attributes/clauses of textual prompts are uniformly complete across all images, neglecting the unavoidable incompleteness of texts/reports in clinical applications and thus making them less feasible. To address this, we, for the first time, identify such incomplete textual prompts in medical image referring segmentation (MIRS) and propose an attribute robust segmentor (ARSeg) by constructing attribute-specific features and balancing the attribute learning procedure. Specifically, based on a U-shaped CNN network and a BERT-based text encoder, an attribute-specific cross-modal interaction module is introduced to establish attribute-specific features, thereby eliminating the dependency of decoding features on complete attributes. To prevent the model from being dominated by attributes with lower missing rates during training, an attribute consistency loss and an attribute imbalance loss are designed for balanced feature learning. Experimental results on two publicly available datasets demonstrate the superiority of ARSeg against SOTA approaches, especially under incomplete and imbalanced textual prompts. Code is available at https://github.com/w7jie/ARSeg.

Keywords: Medical referring image segmentation · Incomplete textual prompts · Imbalanced missing rates.

1 Introduction

Medical image segmentation (MIS), distinguishing lesions or anatomical structures from background, is a crucial yet challenging task in medical image analysis [1,2]. Existing MIS approaches mainly rely on well-annotated medical imaging data for training CNN [3,4], transformer [5], mamba [6], or hybrid architectures [7,8,9]. However, despite their effectiveness, pixel-wisely annotating is expensive [10], making it difficult to collect large-scale medical imaging data to ensure model robustness and generalizability [11,12].

Q. Wang and X. Lin—Equal contribution.

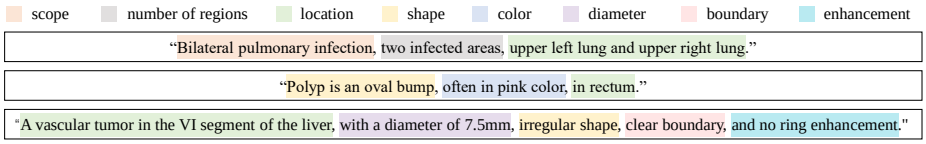

Fig. 1: Examples of textual information. Phrases describing the same attribute are highlighted with the same colors. From top to bottom: textual prompts adopted by LViT [10], MIU-VL [15], and LSMS [13] respectively.

Alternatively, utilizing medical notes or reports as compliments is considered feasible as they are easily accessible and usually contain rich information [10] and thus is promising to promote the convergence and reduce the training costs of MIS models. Therefore, several recent studies have dedicated efforts to investigating language-guided medical image segmentation, firstly defined as medical image referring segmentation (MIRS) by LSMS [13]. Following this, LViT [10] proposes a dual U-shaped architecture to effectively integrate textual and visual information, while Bi-VLGM [14] utilizes a bi-level class-severity-aware vision language graph matching strategy to realize local-class alignment at the word level and global-severity alignment at the sentence level.

Though such pioneering MIRS approaches have demonstrated the feasibility and infinite potential of combining textual information for MIS, the fundamental assumption that textual information is consistently and equally available for each image is unrealistic. On the one hand, collecting complete and uniform texts for each image is of high complexity. On the other hand, relying on complete textual prompts for training and testing may limit the flexibility of MIRS approaches in clinical practice. As depicted in Fig. 1, the textual information of each data source is structured as a fixed set of attributes. In clinical practice, not all predefined attributes are consistently and well recorded. For instance, the predefined attribute set of LViT includes the elements of scope, the number of regions, and locations [10]. In practical applications, it is very often that only scope or number information is available, resulting in incomplete textual prompts compared to the predefined set. Such incomplete textual prompts would result in severe performance degradation in testing and instability in training, which is under-explored.

In this paper, we aim to develop a robust model capable of stably segmenting indicated targets at various attribute completeness. Specifically, we propose a novel framework named ARSeg for attribute-robust MIRS. Based on a U-shaped CNN backbone and a BERT-based text encoder, an attribute-specific cross-modal interaction (ACI) module is developed to capture attribute-specific features for decoding. In this way, the coupling relationship among attributes is released, thereby mitigating the negative impact on feature extraction given missing attributes. To avoid the model being dominated by certain attributes during training especially when textual prompts own imbalanced missing rates, an attribute consistency loss and an attribute balancing loss are jointly utilized.

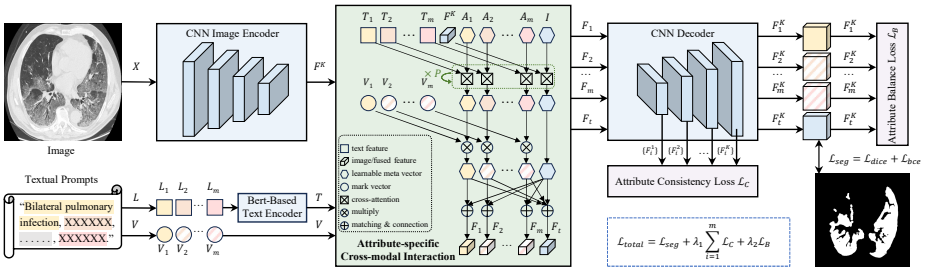

Fig. 2: Overview of the proposed ARSeg.

Experimental results on the QaTa-COV19 and MosMedData+ datasets demonstrate the superiority of ARSeg against existing MIRS approaches on dealing with incomplete and imbalanced textual prompts.

2 Method

The Overall Architecture. As depicted in Fig. 2, ARSeg adopts a partially symmetrical encoder-decoder paradigm. For the encoder part, a progressive downsampling CNN encoder is utilized to extract hierarchical image features, a BERT-based encoder is applied to process the input text prompts for textual feature modeling [16], and an Attribute-specific Cross-modal Interaction(ACI) module is developed to establish the cross-modal interaction between image features and textual features while preserving attribute-specific features. After encoding, each separate existing attribute, together with the whole input text (composed of all existing attributes), will generate individual language-guided multimodal feature maps. For the decoder part, a progressive upsampling CNN decoder is adopted to decode the encoded fused feature maps separately. In addition, an attribute consistency loss and an attribute balancing loss are used to address the interference caused by missing attributes.

Image Encoder and Decoder. The encoder-decoder architecture in ARSeg follows U-Net [3]. Specifically, the CNN encoder is composed of K stacked CNN blocks, and each CNN block consists of a convolution (Conv), a batch normalization (BN), and a ReLu activation layer. To progressively expand the receptive field, a 2×2 max pooling layer is performed after each CNN block. The above encoding procedure is formulated as:

$$F^k = \text{MaxPool}(\text{ReLu}(\text{BN}(\text{Conv}(F^{k-1})))), k \in 1, 2, \ldots, K, \tag{1}$$

where F^k denotes the output of the k-th CNN block. Then, each input image X is encoded as F^K. For decoding, each max pooling in the encoder is replaced with a 2×2 neighboring-based upsampling layer.

Text encoder. Inspired by LViT [10], both BERT and 1D convolution are adapted to process input text prompts. As illustrated in Fig. 1, each text in

MIRS is composed of multiple clauses, each describing an attribute of the target. Given an input text $L = [L_1, L_2, ..., L_m]$ consisting of m clauses/attributes, it is encoded to a series of textual feature embeddings $T = [T_1, T_2, ..., T_m]$. Different from existing MIRS works where all images equally have m attributes, attribute availability of each training image varies in our setting depending on text completeness indicating incomplete textual prompts. To describe the phenomenon of missing attributes, a mark vector $V \in \mathbb{R}^m$ is introduced to indicate the existence of each attribute, where $V_i = 1$ means the i-th attribute exists. To further simulate realistic scenarios where the missing rate of each attribute can be different, a missing rate vector $R \in \mathbb{R}^m$ is defined where $R_i \in [0, 1]$ indicates the missing probability.

Attribute-specific Cross-modal Interaction (ACI). The inputs of ACI consist of an encoded image feature map F^K, a mark vector V generated from the input textual prompts, and encoded textual feature vectors $T = [T_1, T_2, ..., T_m]$. To avoid the excessive coupling among T and between F^K and T when establishing cross-modal interactions, totally $m + 1$ learnable meta vectors are introduced corresponding to the common features of m attributes denoted $\{A_i$, and the input image denoted I. Meanwhile, to enable these meta vectors to learn input-related representations, totally P cross-attention modules are stacked for dependency establishment. Such a process is formulated as:

$$T_i^p = \text{CrossAttention}_p(T_i^{p-1}, A_i^{p-1}, A_i^{p-1}), \tag{2}$$

$$A_i^p = \text{CrossAttention}_p(A_i^{p-1}, T_i^p, T_i^p), \tag{3}$$

where $p = 1, ..., P$ represents the p-th cross-attention module, $i = 1, ..., m$ represents the i-th attribute, given $A_i^0 = A_i$, and $T_i^0 = T_i$. Similarly, the image meta vector I is updated by performing the above interaction procedure starting with F^K. After updating the meta vectors into input-related vectors, matching and connection operations are adopted to generate attribute-specific features for decoding, formulated as:

$$F_i = I^P E^I + V_i A_i^P E_i^T, i = 1, 2, ..., m, \tag{4}$$

where E^I and E_i^T are projection matrices. In addition, the whole-text-guided features F_t for decoding are constructed by fusing the features of all existing attributes with the image features, formulated as:

$$F_t = I^P E^I + \sum_{i=1}^{m} V_i A_i^P E_i^T. \tag{5}$$

Attribute Consistency and Balancing Loss (ACBL). All attribute-specific $\{F_i\}$ and the whole-text-guided features F_t will be fed into the CNN decoder separately, consisting of K CNN blocks for individual decoding. Each block will generate the corresponding decoded features denoted as F_i^k or F_t^k, where $k = 1, ..., K$ represents the k-th decoding block. To achieve accurate prediction of the target under various attribute missing scenarios, an attribute consistency

loss $\mathcal{L}_C$ is introduced to minimize the prediction difference between attribute-specific and the whole-text-guided features, defined as:

$$\mathcal{L}_C^i = \sum_{k=1}^{K} V_i \mathrm{KL}(\sigma(F_i^k), \sigma(F_t^k)), \tag{6}$$

where KL represents the Kullback-Leibler divergence computation and σ represents the Softmax operation.

In clinical applications, attributes are not uniformly missing, with the missing rates of easy-to-obtain attributes being lower and those of others being higher. During training, such imbalanced missing rates will result in different learning speeds across attributes, making the attribute with a lower missing rate dominate the learning direction of the model. To address this, an attribute balancing loss $\mathcal{L}_B$ is designed to supervise the model. Specifically, the prototypes corresponding to the i-th attribute and the whole input text are generated first by averaging F_i^K and F_t^K within the target region, denoted as c_i and c_t. Then, the correlation maps M_i or M_t are generated by calculating the cosine similarity between the prototypes c_i or c_t and the corresponding decoded features F_i^K or F_t^K. Finally, $\mathcal{L}_B$ is formulated as:

$$\eta_i = \frac{\|M_t - M_i\|_2 \sum V}{\sum_{i=1}^{m} V_i \|M_t - M_i\|_2}, \tag{7}$$

$$\mathcal{L}_B = \sum_{i=1}^{m} \eta_i V_i \|M_t - M_i\|_2^2, \tag{8}$$

where $\|\|_2$ represents the L2-norm and η_i reflects the prediction gap between the i-th attribute and the whole text. $\eta_i > 1$ indicates that this attribute has not been well learned. By introducing η_i as weights to penalize prediction differences, $\mathcal{L}_B$ can effectively solve the issue caused by imbalanced missing rates.

For segmentation, a standard segmentation loss $\mathcal{L}_{seg} = \mathcal{L}_{dice} + \mathcal{L}_{bce}$ is adopted, and the overall loss is written as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_1 \sum_{i=1}^{m} \mathcal{L}_C^i + \lambda_2 \mathcal{L}_B, \tag{9}$$

where λ_1 and λ_2 are balancing hyper-parameters.

3 Experiments and Results

Datasets. Two publicly available datasets are selected for evaluation. (1) QaTa-COV19 [17]. It consists of 9,258 chest X-ray images manually annotated with COVID-19 lesions. (2) MosMedData+ [18,19]. It comprises 2,729 CT scan slices manually annotated with lung infections. The data split and textual annotations for QaTa-COV19 and MosMedData+ follow LViT [10]. As illustrated in the first example of Fig. 1, these textual annotations are structured into three attributes,

Table 1: Quantitative comparison in terms of Dice (%) and mIoU (%) on the QaTa-COV19 dataset under different attribute missing rates.

(R_1, R_2, R_3)	LAVT		UniLSeg		RefSegformer		TGANet		LViT		ARSeg	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
(0.0, 0.0, 0.0)	80.48	67.01	72.88	59.58	81.63	69.71	77.17	64.39	81.52	68.63	**84.09**	**72.64**
(0.2, 0.5, 0.8)	75.81	64.32	68.04	54.55	76.30	65.14	75.22	62.95	76.15	65.03	**83.88**	**72.63**
(0.2, 0.8, 0.5)	78.02	67.95	70.02	58.01	77.85	64.71	73.34	60.75	79.32	64.79	**82.80**	**72.51**
(0.5, 0.2, 0.8)	77.95	65.88	69.10	57.82	79.40	68.22	72.02	60.02	76.08	64.13	**80.93**	**68.63**
(0.5, 0.8, 0.2)	78.34	**68.57**	70.08	57.03	77.78	66.35	74.94	62.97	78.97	64.08	**80.10**	67.03
(0.8, 0.2, 0.5)	78.12	66.95	68.67	55.06	78.69	67.63	72.12	59.74	77.27	64.53	**81.71**	**69.52**
(0.8, 0.5, 0.2)	78.25	67.18	71.73	59.84	77.48	65.69	76.07	63.84	78.52	64.97	**80.03**	**67.24**

Table 2: Quantitative comparison in terms of Dice (%) and mIoU (%) on the MosMedData+ dataset under different attribute missing rates.

(R_1, R_2, R_3)	LAVT		UniLSeg		RefSegformer		TGANet		LViT		ARSeg	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
(0.0, 0.0, 0.0)	68.51	55.32	65.89	52.01	70.25	57.31	69.48	55.81	72.10	57.35	**73.24**	**59.82**
(0.1, 0.4, 0.7)	67.08	52.91	63.44	48.29	66.30	52.24	68.07	54.66	69.75	55.06	**72.36**	**58.38**
(0.1, 0.7, 0.4)	66.18	52.24	62.52	47.68	65.69	51.89	66.44	53.31	68.15	55.79	**70.83**	**56.75**
(0.4, 0.1, 0.7)	66.72	52.61	64.12	49.48	68.40	54.35	68.95	55.29	70.19	55.91	**71.74**	**56.15**
(0.4, 0.7, 0.1)	67.05	53.04	60.74	46.04	66.89	53.15	65.49	52.01	70.78	57.08	**72.11**	**57.46**
(0.7, 0.1, 0.7)	64.61	50.57	57.33	43.59	65.72	51.74	67.91	53.94	67.63	53.72	**71.87**	**56.72**
(0.7, 0.4, 0.1)	67.18	53.11	59.61	45.52	66.05	51.99	66.96	56.04	66.38	52.62	**71.38**	**56.64**

including the scope, the number of regions, and locations. Therefore, in our experiments, the number of attributes m is set as 3.

Implementation Details. Models were implemented in PyTorch 1.10.0 and trained on one NVIDIA Geforce RTX 3090 GPU. The Adam optimizer is employed for training, with an initial learning rate of $2e^{-4}$ and a weight decay of $1e^{-4}$. For a fair comparison, all comparison models are re-implemented and trained for 100 epochs on MosMedData+ and 50 epochs on Qata-COV19 under the same settings. To simulate attribute missing, before training, we randomly mask clauses according to imbalanced missing rates R and generate the corresponding mark vector V.

Comparison with State-Of-The-Art. Five SOTA language-guided segmentation models are selected for comparison, including LAVT [20], UnilSeg [21], RefSegformer [22], TGANet [23], and LViT [10]. To evaluate the performance of models under various attribute missing conditions, we test the models under both the complete and six imbalanced attribute missing situations.

Quantitative comparison results on QaTa-COV19 and MosMedData+ are summarized in Tables 1 and 2. Compared to existing approaches, the proposed ARSeg achieves the best Dice and mIoU performance across all attribute missing situations. More importantly, the performance advantage of ARSeg is more pronounced under incomplete textual prompts, achieving the best performance

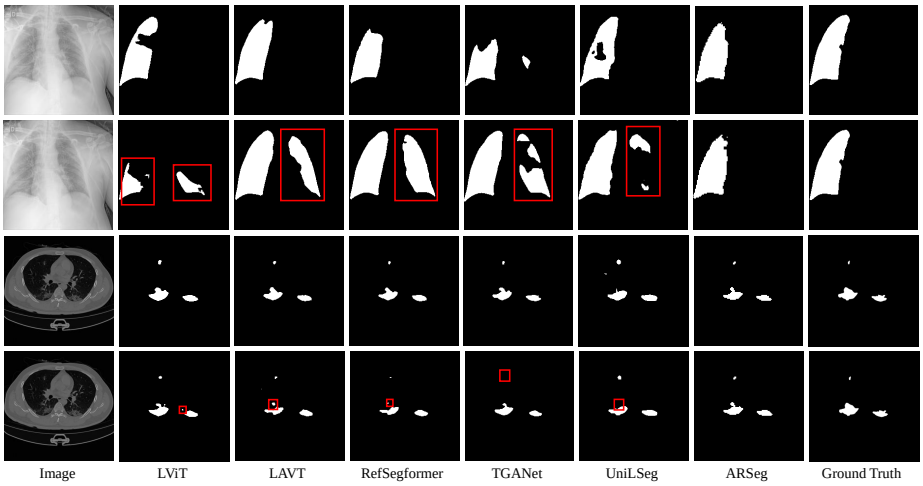

Fig. 3: Exemplar qualitative results of different approaches on QaTa-COV19 (rows 1-2) and MosMedData+ (rows 3-4) under complete (top) and incomplete textual prompts (bottom).

Table 3: Ablation study on different component combinations of ARSeg. Experiments were conducted on QaTa-COV19 with $R = (0.2, 0.5, 0.8)$ and MosMedData+ with $R = (0.1, 0.4, 0.7)$ measured by Dice (%) and mIoU (%).

Components			QaTa-COV19		MosMedData+	
Baseline	ACI	ACBL	Dice	mIoU	Dice	mIoU
●	○	○	80.10	68.48	69.54	54.62
●	●	○	81.39	69.13	71.66	56.23
●	○	●	82.55	71.10	71.29	56.16
●	●	●	**83.88**	**72.63**	**72.36**	**58.38**

stability. Qualitative comparison results are presented in Fig. 3, where all models were trained under complete textual prompts and evaluated under both complete and incomplete textual prompts. Though SOTA approaches accurately identified the target under complete textual prompts, their performance noticeably degrades under incomplete textual prompts. Comparatively, ARSeg maintains outstanding segmentation performance even without complete textual prompts.

Ablation Study on Components. As stated in Table 3, compared to the baseline which completely removes ACI and ACBL module, introducing either ACI or ACBL is helpful, demonstrating their effectiveness for balanced attribute feature learning. After jointly using ACI and ACBL, ARSeg achieves the best performance across both datasets.

Ablation Study on V. To explore the impact of attributes on model performance, we trained and evaluated the models under completely-missing at-

Table 4: Ablation study on V validated on MosMedData+.

Attribute			LViT		Refsegformer		ARSeg	
Scope	Regions	Location	Dice	mIoU	Dice	mIoU	Dice	mIoU
○	○	●	**71.49**	**57.89**	68.56	55.32	70.26	57.09
○	●	○	68.36	53.32	69.94	56.07	**70.38**	**57.23**
●	○	○	70.06	54.33	68.87	54.08	**71.12**	**55.80**
○	●	●	69.97	56.28	66.61	52.64	**70.98**	**56.91**
●	●	○	69.64	55.09	66.96	52.60	**71.59**	**56.57**
●	○	●	**71.32**	**58.04**	68.65	53.84	70.54	55.04
●	●	●	72.10	57.65	70.25	57.31	**73.24**	**59.82**

Table 5: Ablation study on P validated on QaTa-COV19 dataset with $R = (0.2, 0.5, 0.8)$ and MosMedData+ dataset with $R = (0.1, 0.4, 0.7)$.

P	QaTa-COV19		MosMedData+	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
3	82.85	69.91	70.58	55.27
5	**83.88**	**72.63**	**72.36**	**58.38**
7	81.53	69.27	71.49	56.24

tributes. As summarized in Table 4, under one attribute prompting, using Location as textual prompts achieves the best performance. As the number of attributes increases, textual information becomes richer and the model performance gradually improves. Compared to SOTA approaches, ARSeg achieves the best performance under five out of seven settings and is much more stable against missing attributes.

Ablation Study on *P*. Quantitative results under various settings of P are summarized in Table 5. When using fewer cross-attention modules (*i.e.*, a smaller P), the meta vectors may not be fully updated into input-related vectors, resulting in suboptimal prediction results. Comparatively, given a larger P, elements within each meta vector may constantly interact with the same input features, leading to poorer feature diversity. Therefore, the selection of P is task-dependent and setting $P = 5$ achieves the best performance on QaTa-COV19 and MosMedData+.

4 Conclusion

In this paper, we extend the MIRS task into a more realistic and challenging scenario, namely language-guided segmentation with incomplete textual prompts. To address this issue, we propose ARSeg, a vision-language model for balanced attribute feature learning under imbalanced attribute distributions in textual prompts. Specifically, based on the U-shaped CNN backbone and the BERT-based encoder, an attribute-specific cross-modal interaction module is proposed

to establish attribute-specific features for decoding, which allows the decoder to accurately decode the features driven by any single attribute. In addition, an attribute consistency loss and an attribute balancing loss are developed to stabilize the training procedure under imbalanced attribute missing rates. Experimental results across various attribute missing rates on two publicly-available datasets validate the effectiveness of ARSeg for robust language-guided medical image segmentation.

Acknowledgment This work was supported in part by the National Natural Science Foundation of China under Grants 62202179 and 62271220, and in part by the Foundation Research Fund of Shenzhen (2024534319).

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

References

1. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods. **18**(2), 2023–2011 (2021)
2. Zhou, H. Y., et al.: nnFormer: Volumetric medical image segmentation via a 3D transformer. IEEE Trans. Image Process. **42**, 4036–4045 (2023)
3. 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
4. Chen, G., Li, L., Dai, Y., Zhang, J., Yap, M. H.: AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. IEEE Trans. Med. Imag. **42**(5), 1289–1300 (2023)
5. Li, J., et al.: Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. **85**, 102672 (2023)
6. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, Lei.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: Linguraru, M. G., et al. (eds.) MICCAI 2024, LNCS, vol. 15008, pp. 578–588. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72111-3_54
7. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Trans. Med. Imag. **42**(9), 2763–2775 (2023)
8. Roy, S., et al.: Mednext: transformer-driven scaling of convnets for medical image segmentation.. In: Greenspan, H., et al. (eds.) MICCAI 2023, LNCS, vol. 14223, pp. 405–415. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_39
9. Guo, X., Lin, X., Yang, X., Yu, L., Cheng, K. T., Yan, Z.: UCTNet: Uncertainty-guided CNN-Transformer hybrid networks for medical image segmentation. Pattern Recogn. **152**, 110491 (2024)

10. Li, Z., et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE Trans. Med. Imag. **43**(1), 96–107 (2023)
11. Lin, X., Wang, Z., Yan, Z., Yu, L.: Revisiting Self-attention in Medical Transformers via Dependency Sparsification. In: Linguraru, M. G., et al. (eds.) MICCAI 2024, LNCS, vol. 15011, pp. 555–566. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_52
12. Azad, R., et al.: Medical image segmentation review: The success of u-net. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10076–10095 (2024)
13. Ouyang, S., et al.: LSMS: Language-guided scale-aware medsegmentor for medical image referring segmentation. arXiv preprint arXiv:2408.17347 (2024)
14. Chen, W., Liu, J., Liu, T., Yuan, Y.: Bi-VLGM: Bi-level class-severity-aware vision-language graph matching for text guided medical image segmentation. Int. J. of Comput. Vision. 1–17 (2024)
15. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. In: International Conference on Learning Representations (2023)
16. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)
17. Degerli, A., Kiranyaz, S., Chowdhury, M. E. H., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: IEEE International Conference on Image Processing, pp. 2306–2310 (2022)
18. Morozov, S. P., et al.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
19. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Experim. **4**(1), 1–13 (2020)
20. Yang, Z., et al.: Language-aware vision transformer for referring segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 1–18 (2024)
21. Liu, Y., Zhang, C., Wang, Y., Wang, J., Yang, Y., Tang, Y.: Universal segmentation at arbitrary granularity with language instruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3459–3469 (2024)
22. Wu, J., Li, X., Li, X., Ding, H., Tong, Y., Tao, D.: Towards robust referring image segmentation. IEEE Trans. Image Process. **33**, 1782–1794 (2024)
23. Tomar, N. K., Jha, D., Bagci, U., Ali, S.: TGANet: Text-guided attention for improved polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, LNCS, vol. 13433, pp. 151–160. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_15