Enabling Text-free Inference in Language-guided Segmentation of Chest X-rays via Self-guidance

Shuchang Ye[1 [0009−0006−1935−1953]], Mingyuan Meng[1,2 [0000−0002−9562−1613]], Mingjian Li[1 [0009−0001−5591−5385]], Dagan Feng[1 [0000−0002−3381−214X]], and Jinman Kim[1 (✉) [0000−0001−5960−1060]]

[1] The University of Sydney, Sydney NSW 2000, Australia
jinman.kim@sydney.edu.au
[2] Shanghai Jiao Tong University, Shanghai 200240, China

Abstract. Segmentation of infected areas in chest X-rays is pivotal for facilitating the accurate delineation of pulmonary structures and pathological anomalies. Recently, multi-modal language-guided image segmentation methods have emerged as a promising solution for chest X-rays where the clinical text reports, depicting the assessment of the images, are used as guidance. Nevertheless, existing language-guided methods require clinical reports alongside the images, and hence, they are not applicable for use in image segmentation in a decision support context, but rather limited to retrospective image analysis after clinical reporting has been completed. In this study, we propose a self-guided segmentation framework (SGSeg) [3] that leverages language guidance for training (multi-modal) while enabling text-free inference (uni-modal), which is the first that enables text-free inference in language-guided segmentation. We exploit the critical location information of both pulmonary and pathological structures depicted in the text reports and introduce a novel localization-enhanced report generation (LERG) module to generate clinical reports for self-guidance. Our LERG integrates an object detector and a location-based attention aggregator, weakly-supervised by a location-aware pseudo-label extraction module. Extensive experiments on a well-benchmarked QaTa-COV19 dataset demonstrate that our SGSeg achieved superior performance than existing uni-modal segmentation methods and closely matched the state-of-the-art performance of multi-modal language-guided segmentation methods.

Keywords: Language-guided Segmentation · Multi-Modal Learning

1 Introduction

Chest X-rays play an essential role in the diagnosis of some pulmonary infectious diseases. In the analysis of chest X-rays, segmentation of the infected areas is essential for improving diagnostic accuracy, optimizing treatment plans, and

[3] the code repository can be accessed at https://github.com/ShuchangYe-bib/SGSeg

enabling disease progression monitoring [1]. However, manual segmentation conducted by radiologists is labor-intensive and prone to inconsistencies, posing challenges to its scalability and uniformity in clinical applications [2]. This inspires the integration of deep learning into the segmentation of chest X-rays, offering a pathway to automate the delineation process and enhance the efficiency and reliability of pulmonary diagnosis [3, 4]. The development of medical segmentation has been significantly advanced since the invention of U-Net [5, 6]. With the progression of neural network architectures, U-Net has been extended as many variants (e.g., U-Net++ [7], Attention U-Net [8], Trans U-Net [9], and Swin U-Net [10]) and obtained improved performance. Nevertheless, a persistent challenge within the medical domain remains: the inherent complexity of medical images poses difficulties for models to interpret underlying pathologies and identify disease locations, resulting in suboptimal segmentation accuracy for pulmonary lesions.

Recently, multi-modal learning has provided evidence that integrating visual and textual data exhibits superior performance over their uni-modal counterparts [11]. Visual Language Pre-training (VLP) [12] has significantly advanced across various computer vision tasks by effectively bridging image and text features. For instance, CLIP [13] adopted contrastive learning techniques to align the representations of image and text in latent space, fostering robust cross-modal similarities. Existing VLP primarily trained encoders, yet for tasks requiring both an encoder and a decoder, such as segmentation, a more comprehensive training approach to simultaneously optimize both components is essential. In medical image segmentation, LViT [14] demonstrated that the models guided by additional textual information can achieve higher performance. Building upon the LViT, Zhong et al. [15] advanced image-text fusion techniques by introducing a text-guided decoder in U-Net. These multi-modal language-guided methods outperformed existing uni-modal methods, achieving state-of-the-art performance in chest X-ray segmentation. However, existing multi-modal methods necessitate the input of textual reports alongside images during inference, diverging from the clinical protocol of analyzing images prior to generating reports and thus reducing their clinical applications.

In this study, we explore leveraging linguistic context during training while enabling text-free inference in language-guided segmentation of chest X-rays. Our main contributions are summarized as follows:

- We propose a self-guided segmentation framework (SGSeg) where the encoder-decoder process is self-guided by generated clinical reports during inference.
- Our SGSeg introduces a novel Localization-Enhanced Report Generation (LERG) module that can accurately identify disease locations and generate reports to provide guidance for segmentation by utilizing object predictions from an object detector to assist the report generation process.
- To address the issue where most object predictions produced by object decoders are categorized as “no class," with only a minor portion accurately indicating the locations of infected areas, we proposed a location-based attention aggregator to transform sparse object prediction into location features.

– To provide weak supervision of LERG, a clustering-based location-aware pseudo-label extractor is introduced to extract location information from clinical reports.

Extensive experiments on the well-benchmarked QaTa-COV19 dataset demonstrate that the proposed SGSeg outperforms existing uni-modal inference segmentation methods and closely approximates the benchmarks set by the state-of-the-art multi-modal inference segmentation methods.

2 Method

2.1 Self-guided segmentation framework (SGSeg)

The proposed Self-Guided Segmentation framework (SGSeg) comprises two main components: a Language-guided U-Net and a novel weakly-supervised localization-enhanced report generation (LERG) component (see Fig. 1). During the training phase, ground-truth reports served as inputs to the text encoder from which labels were extracted to provide weak supervision for LERG. In the inference phase, generated reports replaced the ground truth as inputs to the text encoder, facilitating text-free inference.

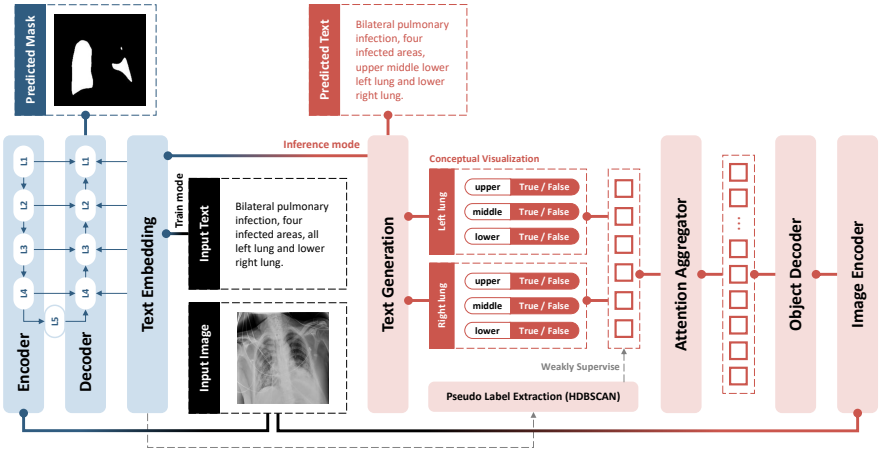

Fig. 1. The neural network architecture of the proposed SGSeg framework: The blue segment represents the Language-guided U-Net, while the pink segment denotes localization-enhanced report generation processes.

2.2 Language-guided U-Net

In the Language-guided U-Net, the image downsampling process utilized ConvNeXt-T [17], pre-trained on ImageNet-1K, to sequentially reduce the dimen-

sionality from 224x224 to 7x7 by a factor of 4 at each step. For the image upsampling decoder, we adopted the GuidedDecoder structure from LanGuideMedSeg [15], utilizing a cross-modal attention module to fuse extracted position information with image features effectively. The text encoder was implemented using BERT [19], which was pre-trained through masked language modeling [18] and multimodal contrastive learning [13] on the MIMIC dataset.

2.3 Localization-enhanced Report Generation

Understanding Text's Role in Language-guided Segmentation To delve into the underlying principles of how text influences the segmentation results, we examined the importance of each word within a given text through a cross-modal attention module, as shown in Fig. 2. The importance of each word was estimated based on the product of the query (q) and key (k) vectors for each word. The heatmap revealed a pronounced emphasis on location-descriptive words ("upper", "middle", "lower", "left", and "right") within the cross-modal attention framework. This phenomenon indicated the potential to refine segmentation accuracy by localizing infected areas.

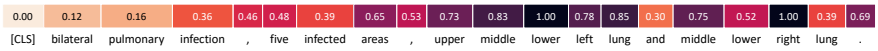

Fig. 2. Illustration of word importance, where we visualize the attention weight on each token of a report.

Location-aware pseudo-Label Extraction To effectively use the additional textual report during training, we first extract descriptions related to location information, utilizing a BERT model trained on X-ray to encode disease location into latent representation. Subsequently, we apply HDBSCAN [20] to group the text embedding into meaningful clusters that reflect the spatial relationships among reports.

Weakly-supervised Localization-enhanced Report Generation We propose automatically generating reports focusing on spatial positioning to guide segmentation. Given the absence of ground truth labels for location prediction, pseudo-labels were extracted from reports to weakly-supervise the localization process. Our object detection network adhered to the RT-DETR [21] architecture, where images were first compressed via a CNN architecture, followed by intra-scale feature interaction through self-attention. Subsequently, features of varying granularity interact via a Cross-Scale Feature-Fusion Module (CCFM), with which object queries are decoded into object predictions via an object decoder, according to:

$$Q = Deocde(CCFM(F_{CNN}(I))) \tag{1}$$

where P represents object prediction derived from the image I processed through ResNet50 [22] backbone and the CCFM transformation. To refine the alignment between the predicted vector, denoted as p, and the pseudo-labels, represented by y, the Binary Cross-Entropy Loss was employed according to:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \tag{2}$$

where N signifies the number of labels, specifically six for this framework, aligning with the "upper", "middle", and "lower" regions across both lungs. Here, p_i and y_i denote the predicted probability and the actual label for the $i-th$ label, respectively.

The final step involves decoding the labels into precise infected area locations, enabling the inference of both the number of infected areas and the overall infection status across the lungs. This decoded information is then synthesized into a text description.

Location-based Attention Aggregation This module is designed to refine sparse object predictions from the object decoder into location features by initializing a location-specific query vector q. The process involves calculating attention weights for each object prediction through matrix multiplication. Subsequently, we derive the aggregated features $A = softmax(Xq^T) \cdot X$ by the linear combination of these weighted object predictions, where X denotes the input object predictions.

3 Experiments

3.1 Dataset

The dataset used to evaluate our methodology was the QaTa-COV19 dataset [16] [4], the only publicly available chest X-ray dataset with text medical reports. It comprises $9,258$ chest X-ray images of COVID-19 infections alongside segmentation annotations of corresponding infection regions. This dataset was augmented by Li et al. by providing brief, structured, textual descriptions detailing the infection site [14]. The dataset was partitioned adhering to the official split [15] into $5,716$ for training, $1,429$ for validation, and $2,113$ for testing.

Subsequent refinement of the dataset [5] involved the correction of erroneous descriptions, typographical errors, and ambiguous expressions, which impacted 4% of the data. Pseudo-labels were derived from textual annotations to provide weak supervision of lesion localization.

[4] The Qata-COVID-19 dataset used in this study can be accessed at the following URL: https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset.

[5] The refined dataset is available at https://github.com/ShuchangYebib/SGSeg/tree/main/data

Table 1. Performance comparison between our SGSeg and existing uni-modal and multi-modal segmentation methods on the QaTa-COV19 dataset. The best results of uni- and multi-modal methods are underlined. The results of our SGSeg are highlighted in bold.

Modality	Model	Accuracy	Dice	Jaccard
Uni-Modal	U-Net[5]	0.945	0.819	0.692
	U-Net++[7]	0.947	0.823	0.706
	Attention U-Net[8]	0.945	0.822	0.701
	Trans U-Net[9]	0.939	0.806	0.687
	Swin U-Net[10]	<u>0.950</u>	<u>0.832</u>	<u>0.724</u>
Multi-Modal Train, Uni-Modal Inference	SGSeg (ours)	**0.971**	**0.874**	**0.778**
Multi-Modal	LViT[14]	0.962	0.837	0.751
	LanGuideSeg[15]	<u>0.975</u>	<u>0.898</u>	<u>0.815</u>

3.2 Implementation Details

We used the image size of 224×224 with a hidden dimension 768 in the cross-attention module. PyTorch [24] and PyTorch Lightning [25] were used as the development environment with NVIDIA RTX A6000 GPUs. For training, we applied AdamW optimizer [26] with a cosine annealing learning rate policy (initial rate 3×10^{-4}) and reduced it to $< 1 \times 10^{-6}$. The batch size was set to 32. Data augmentations, including random crops, masks, and rotations, were applied.

4 Result and Discussion

4.1 Comparison with Existing Methods

To evaluate the SGSeg framework's efficacy, comparative analyses were conducted with current uni-modal and multi-modal segmentation models, as shown in table 1. Results illustrate that SGSeg exceeds the performance of conventional uni-modal methods and closely matches that of advanced multi-modal approaches. Relative to the leading uni-modal inference model, our approach achieved a notable enhancement in accuracy from 0.950 to 0.971 (2.21%), in the Dice coefficient from 0.832 to 0.874 (5.05%), and in the Jaccard index from 0.724 to 0.778 (7.46%). However, when compared with the top-performing multi-modal inference model, our method exhibited a slight decrease in accuracy by 0.41%, and reductions in the Dice coefficient and Jaccard index by 2.74% and 4.75%, respectively. Fig. 3 illustrated the significant impact of textual information on enhancing segmentation accuracy, particularly in challenging cases. It demonstrated that incorporating location-specific or pseudo-location data facilitates the model's precision in identifying pathological areas. The SGSeg model was trained to identify lesions through weak supervision provided by additional textual data, compensating for the absence of textual descriptions at inference

by autonomously generating this essential information. Consequently, the model leveraged guidance from the detector to achieve enhanced segmentation outcomes.

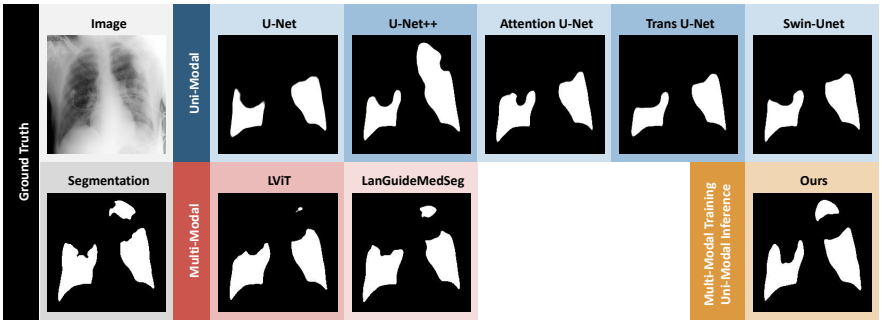

Fig. 3. Comparative Analysis of Segmentation Results: Uni-modal vs. Multi-modal Methods

4.2 Ablation Study

Table 2. Ablation studies on the impact of individual components in SGSeg: Without text - uni-modal segmentation; Visual-language Pre-training - pre-trained with CLIP and fine-tuned by uni-modal segmentation; Self-Guidance - using generated text as input during inference; Full Text - using ground truth text as input during inference.

	Accuracy	Dice	Jaccard
Without Text	0.953	0.846	0.725
Vision-language pre-training (CLIP)	0.962	0.850	0.740
Self-Guidance (simple report generation)	0.966	0.861	0.759
Self-Guidance (weakly-supervised LERG)	0.971	0.874	0.778
Full Text	0.973	0.890	0.797

Our framework leverages additional language information during training while eliminating the need for textual input during inference. To demonstrate the utility of synthetic text and validate our self-guided design, we conducted an ablation study (Table 2). The results show that including additional textual information significantly improves segmentation performance. Comparing vision-language pre-training with our method, our multi-modal encoder-decoder training outperformed multi-modal encoder-focused pre-training. SGSeg significantly outperformed methods without textual input and closely matched those using ground truth text for inference. This indicates that our self-guidance framework

effectively utilizes text for weak supervision during training and autonomously generates text inputs for inference. Additionally, the proposed weakly supervised LERG module enhanced segmentation accuracy, underscoring the efficacy of incorporating a location-aware pseudo-label extractor and a location-based attention aggregator.

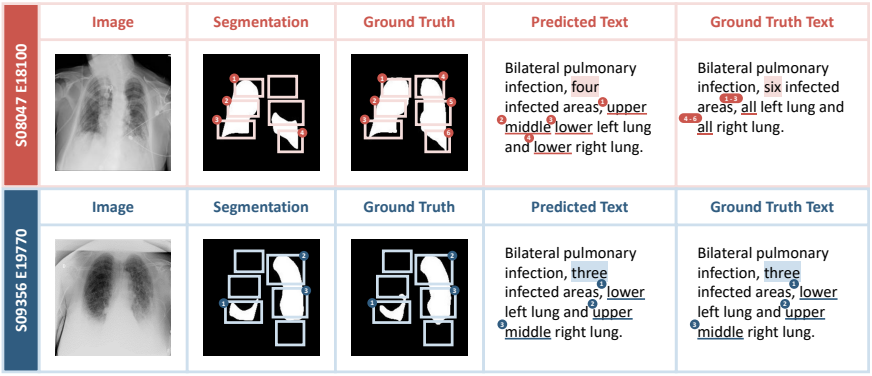

Fig. 4. Comparative analysis on the relation between generated text and segmentation outcomes

4.3 Visualization

We conducted an in-depth examination of the attention maps and the relationship between the generated text and the segmentation outcomes. Our model demonstrates accurate attention to lesions, which is fundamental to segmentation (see Fig. 5). Notably, the transition from exposure to ground truth text during training to reliance on generated text during inference impacts segmentation outcomes. Comparative analysis, as depicted in Fig. 4, indicates that inaccuracies in generated reports moderately influence the model's performance.

5 Conclusion

This study identified a crucial shortfall in current language-guided segmentation methods: their reliance on textual inputs during inference diminishes their relevance in clinical practice. To overcome this challenge, we analyzed the text's role in language-guided segmentation and proposed an innovative self-guided segmentation framework tailored for text-free analysis. Experiments on the QaTa-COV19 dataset showed that our SGSeg significantly outperformed existing unimodal image-only methods and closely approached the multi-modal methods requiring text reports during inference.

Fig. 5. Visualization of the model's attention distribution over the input image arranged sequentially as follows: image, ground truth segmentation, attention map, and attention map projected onto the image

Limitation A limitation of this study is its reliance on the QaTa dataset for experiments. While the QaTa dataset is comprehensive and widely recognized for language-guided segmentation tasks, the findings may not fully generalize to other imaging modalities or datasets.

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

References

1. J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, pp. 654, 2024, Nature Publishing Group UK London.
2. Nillmani, N. Sharma, L. Saba, N. N. Khanna, M. K. Kalra, M. M. Fouda, and J. S. Suri, "Segmentation-Based Classification Deep Learning Model Embedded with Explainable AI for COVID-19 Detection in Chest X-ray Scans," *Diagnostics*, vol. 12, no. 9, article 2132, 2022.
3. T. Mahmood, A. Rehman, T. Saba, L. Nadeem, and S. A. O. Bahaj, "Recent advancements and future prospects in active deep learning for medical image segmentation and classification," *IEEE Access*, 2023, IEEE.
4. S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021, Springer.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, 2015, Springer.
6. N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021, IEEE.
7. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3–11, 2018, Springer.

8. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
9. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
10. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, pp. 205–218, 2022, Springer.
11. Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multimodal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10944–10956, 2021.
12. Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, et al., "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations and Trends®in Computer Graphics and Vision*, vol. 14, no. 3–4, pp. 163–352, 2022, Now Publishers, Inc.
13. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, 2021, PMLR.
14. Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023, IEEE.
15. Y. Zhong, M. Xu, K. Liang, K. Chen, and M. Wu, "Ariadne's Thread: Using Text Prompts to Improve Segmentation of Infected Areas from Chest X-ray Images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 724–733, 2023, Springer.
16. A. Degerli, S. Kiranyaz, M. E. H. Chowdhury, and M. Gabbouj, "Osegnet: Operational Segmentation Network for Covid-19 Detection Using Chest X-Ray Images," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2306-2310, 2022, doi: 10.1109/ICIP46576.2022.9897412.
17. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.
18. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
19. B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing," *arXiv preprint arXiv:2204.09817*, 2022. https://arxiv.org/abs/2204.09817
20. R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 160–172, 2013, Springer.
21. W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs Beat YOLOs on Real-time Object Detection," *arXiv preprint arXiv:2304.08069*, 2023.
22. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

23. J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 92–100, 2019, Springer.
24. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
25. W. A. Falcon, "Pytorch lightning," *GitHub*, vol. 3, 2019.
26. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.