# Common Vision-Language Attention for Text-Guided Medical Image Segmentation of Pneumonia

Yunpeng Guo, Xinyi Zeng, Pinxian Zeng, Yuchen Fei, Lu Wen, Jiliu Zhou, Yan Wang✉

School of Computer Science, Sichuan University, Chengdu, China
wangyanscu@hotmail.com

**Abstract.** Pneumonia, recognized as a severe respiratory disease, has attracted widespread attention in the wake of the COVID-19 pandemic, underscoring the critical need for precise diagnosis and effective treatment. Despite significant advancements in the automatic segmentation of lung infection areas using medical imaging, most current approaches rely solely on a large quantity of high-quality images for training, which is not practical in clinical settings. Moreover, the unimodal attention mechanisms adopted in conventional vision-language models encounter challenges in effectively preserving and integrating information across modalities. To alleviate these problems, we introduce Text-Guided Common Attention Model (TGCAM), a novel method for text-guided medical image segmentation of pneumonia. Text-Guided means inputting both an image and its corresponding text into the model simultaneously to obtain segmentation results. Specifically, TGCAM encompasses the introduction of Common Attention, a multimodal interaction paradigm between vision and language, applied during the decoding phase. In addition, we present an Iterative Text Enhancement Module that facilitates the progressive refinement of text, thereby augmenting multi-modal interactions. Experiments respectively on public CT and X-ray datasets demonstrated our method outperforms the state-of-the-art methods qualitatively and quantitatively.

**Keywords:** Medical Image Segmentation · Multi-Modal · Common Attention

## 1 Introduction

Pneumonia, a severe respiratory illness that affects the alveoli and distal airways, presents significant health risks [1]. The persistent threat of the coronavirus disease 2019 (COVID-19) pandemic [2] has heightened concerns regarding the diagnosis and treatment of pneumonia. Traditionally, the diagnostic process for pneumonia has heavily relied on the manual interpretation of various radiological imaging modalities, demanding considerable time and expertise from radiologists. Thanks to the development of deep learning, numerous methodologies grounded in various types of neural networks, such as convolutional neural networks (CNNs) [3-6] and transformers, have been proposed and successfully applied in

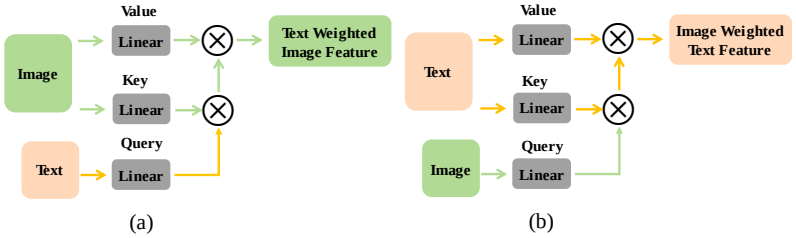


**Fig. 1.** Two kinds of unimodal attention mechanisms

medical image segmentation tasks [7-10]. These approaches facilitate the precise delineation of the infected areas, significantly reducing the reliance on extensive manual effort and domain-specific specialized knowledge.

In the field of medical image segmentation, existing methods predominantly rely on large amounts of labeled images for training, imposing high demands on the quality and quantity of available data. To mitigate similar issues available in natural image processing, CLIP [11] capitalizes on the complementary information provided by accompanying text, thereby reducing the dependency on high-quality annotated image data and maximizing the utilization of available information. Inspired by CLIP, approaches like MedCLIP [12] and GLoRIA [13] have extended these ideas to medical image-report pairs situations, still showcasing excellent performance. Motivated by the impressive performance gains achieved through the integration of textual information, Li et al. [14] proposed a novel CNN-Transformer structure named LViT to integrate multimodal information in the early stage. Lee et al. [15] propose a Text-Guided Cross-Position attention module, eschewing the commonly used transformer architecture. Zhong et al. [16] establish a more robust baseline using a U-Net variant architecture and strategically incorporate multimodal fusion processes into the decoding stage, leading to state-of-the-art performance enhancements.

While the aforementioned studies have introduced text information, they apply the typical cross attention to obtain the Enhanced image feature by updating the value of the image with the query of text as shown in Fig. 1(a) [15,16]. This process represents image features using text-weighted attention. Conversely, Fig. 1(b) represents text features using image-weighted attention. These two forms of cross-attention, which we term unimodal attention mechanisms, only capture a portion of multimodal information, and information of the model acting as the query is not fully preserved. To alleviate this issue, we propose common attention, a novel attention mechanism to obtain genuine multimodal features. By integrating two unimodal attentions, our common attention mechanism facilitates deep and equitable interactions between visual and textual information, thereby enhancing the preservation and understanding of multimodal information. Moreover, unlike approaches that either introduce text features only once at the input stage [14,15,17] or utilize the same features obtained after initial encoding at each fusion stage [16], our method prevents discrepancies in feature levels between modalities, ensuring that image features correspond

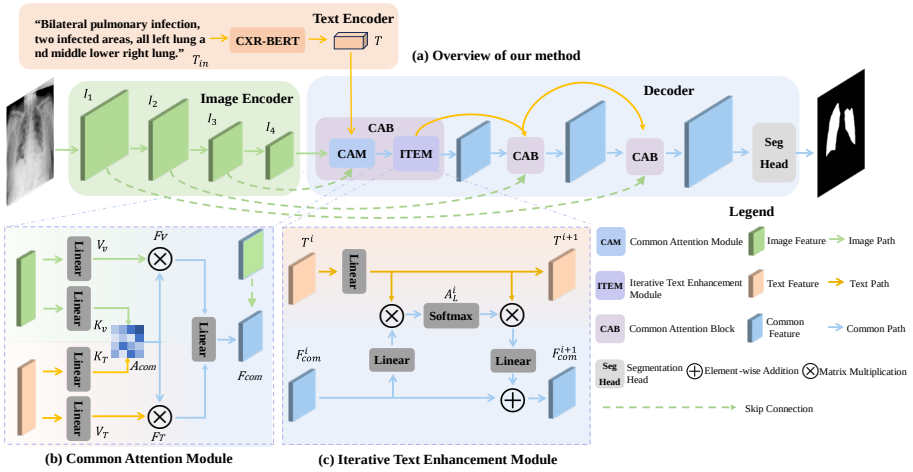


**Fig. 2.** Overview of our proposed method and detail of the modules. Our model mainly consists of an image encoder, a text encoder and a multi-modal decoder. The decoder contains three Common Attention Blocks (CAB) to fuse and decode the multimodal features layer-by-layer, while each CAB contains a Common Attention Module (CAM) and Iterative Text Enhancement Module (ITEM).

consistently to the same level of text features. To achieve this goal, we propose an Iterative Text Enhancement Module, which iteratively converts text features to maximize their fusion capabilities. Our contributions can be summarized as follows. 1. We propose TGCAM, a text-guided segmentation method that incorporates text information for the automatic segmentation of infected regions, yielding excellent performance results. 2. We introduce a multimodal interaction mechanism named common attention and additionally design an Iterative Text Enhancement Module to facilitate deeper interaction between text and images. 3. We conduct experiments on two public datasets, QaTa-Cov19 [18] and MosMedData [19], achieving a new state-of-the-art and demonstrating the effectiveness of the proposed method and individual modules.

## 2 Method

In this section, we innovatively propose a Text-Guided Common Attention Model (TGCAM). The overall architecture of our method is illustrated in Fig. 2. The model consists of three main components: text encoder, image encoder, and decoder. The image encoder encodes the input image to obtain image features, while the text encoder encodes the corresponding clinical text information to obtain text features. Subsequently, both the text and image features are passed into the Common Attention Block (CAB) for effective

fusion of the multimodality information. Specifically, they first pass the Common Attention Module, following the proposed common attention mechanism, which attends to preserve the intrinsic information of two modalities, to obtain fused multimodal features. Then, taking the fused multimodality features and the text feature of this layer as input, the Iterative Text Enhancement Module converts the text features to different levels iteratively, meanwhile promoting the deeper interaction of multimodal features. Repeated iterations of the CAB module better utilize image and text features at different levels, resulting in more effective fusion representations. Finally, the Segmentation Head is used to obtain the prediction. This architecture is designed with low coupling and portability, separating the encoding and fusion steps.

### 2.1 Encoder Design

Following [16], we adopt the ConvNeXt-Tiny [20] as the structure of the image encoder. It gradually extracts image features at different levels during the encoding phase: $I_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_1}$, $I_2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$, $I_3 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_3}$, $I_4 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C_4}$. Note that, $H$ and $W$ represent the height and width of the input image, $C_i$ is the number of channels. For the Text Encoder, we employ CXR-BERT [21] to faithfully capture the text features. Taking the clinical text information $T_{in}$ as the text input, the text encoder outputs the text feature $T \in \mathbb{R}^{N\times C}$, where $N$ denotes the length of the text feature and $C$ represents text channels.

### 2.2 Common Attention Module

Current widely adopted unimodal attention, respectively transformed the features of the image and text into Key and Query and then calculate the attention matrix to update the image one. In this manner, the textual feature could provide crucial information for the visual feature to learn the importance of the region described by text information but doesn't participate as the output to the next layer, which leads to the loss of pure text information. To alleviate this issue, we design a Common Attention Module which drops the primary and secondary relationship in conventional attention mechanism, thus effectively enhancing the interaction between text and image. Specifically, as shown in Fig. 2(b), the proposed Common Attention Module takes text and image features as input, projecting them into Keys and Values without Query by different linear layers, treating inputs from both modalities equally, instead of considering one modality as a query condition for another modality, and calculates the attention matrix $A_{com} \in \mathbb{R}^{N\times HW}$ with the Key of each modality, taking both as the subject information to be retained, to obtain the common attention by Eq.1, which represents the equal relationship between them. where $\frac{1}{\sqrt{c}}$ is the scaling factor.

$$A_{com} = \frac{1}{\sqrt{c}} K_T (K_V)^T. \tag{1}$$

After softmax normalization in different directions, $A_{com}$ is used to compute $F_V$ and $F_T$ as follows:

$$F_V = softmax(A_{com})V_v, \tag{2}$$
$$F_T = softmax\left(A_{com}{}^T\right)V_T. \tag{3}$$

The obtained features $F_V \in \mathbb{R}^{N\times C}$ and $F_T \in \mathbb{R}^{HW\times C}$ represents the weighted image feature attended by each word in the sentence and the weighted textual feature attended by each pixel in the image, respectively. Finally, we fuse these two kinds of features into the genuine multimodal feature, that is, the common feature $F_{com} \in \mathbb{R}^{HW\times N}$, which is computed as shown in Eq.4.

$$F_{com} = F_T(F_V)^T. \tag{4}$$

$F_{com}$ is then reshaped to the size of $HW \times C$ by linear projection function and passed to the next Module.

### 2.3 Iterative Text Enhanced Module

Generally, text features are incorporated only once during the decoding stage, while image features are introduced with the skip connection repeatedly. Nevertheless, this may result in the modality imbalance from the genuine true multimodal perspective. Therefore, we also intend to introduce the textual information in multiple layers with reference to [16]. However, it directly introduces the same text features obtained after initial text encoding at each fusion stage. Actually, due to the introduction of different levels of image features from the skip connection, the text features of interest should be different for different layers. That is to say, high-level visual features should merge with high-level textual features, and the same applies to low-level visual and textual features. Thus, we propose the Iterative Text Enhancement Module, which iteratively converts text features, realizing better deep inter-modal interaction. As shown in the Fig. 2(c), the Iterative Text Enhancement Module is behind the Common Attention Module. It takes the text feature $T$ and the fused feature $F_{com}$ of each layer as input. For the first layer, the text features derive from the initial encoding. The text features $T$ and fused feature $F_{com}$, after linear projection, are used to obtain the intermediate attention matrix $A_L$ as follows, where $i$ represents the number of layers and $LP$ [.] represents the linear projection function.

$$T^{i+1} = LP[T^i], \tag{5}$$
$$A_L^i = softmax(LP[F^{\,i}_{com}](T^{i+1})^T). \tag{6}$$

The intermediate feature is obtained through the attention matrix, and another linear projection function is performed. The result is summed with $F^{\,i}_{com}$ to get $F^{\,i+1}_{com}$, while BN represents batch normalization. Finally, the new common features $F^{\,i+1}_{com}$ and text features $T^{i+1}$ are fed into the next layer as input image features and text features:

$$F^{\,i+1}_{com} = BN\left(F^{\,i}_{com} + LP\left[A_L^i T^{i+1}\right]\right) \tag{7}$$

### 2.4 Decoder Module

The Decoder Module includes three Common Attention Blocks. In the first block, initial image and text features are input. After the Common Attention Block, a skip connection is

performed, followed by the up-sampling. The processed common feature replaces the pure visual features in the first block and is sent to the next block with iteratively converted text features. After all fusion blocks, the final multimodal information is fed into the Segmentation Head to obtain the predicted results.

# 3 Experiments

## 3.1 Datasets

We used two datasets, QaTa-COV19 [18] and MosMedData [19], to evaluate the performance of our proposed method. Both datasets are medical image datasets derived from COVID-19, containing image sets of X-ray and CT modalities, respectively. QaTa-COV19 consists of 9258 chest X-ray images from COVID-19 patients, collected by researchers from Qatar University and Tampere University. Specifically, 7145 are in the train sets, and 2113 are in the test sets. The dataset also includes annotated lung infection regions as ground truths for segmentation tasks. MosMedData is a large lung CT scan dataset for COVID-19, compiled from seven public datasets. There are a total of 2729 pairs of images and matching ground truth masks. To ensure consistency between different datasets, the annotations for different types of lesions are all colored white.

It's worth noting that both datasets originally only contained image data. Li et al. [14], with the assistance of experts, manually annotated text information for both datasets. For example, shown in Fig. 2, "Bilateral pulmonary infection, two infected areas, upper left lung and upper right lung" focuses on whether there are infections on both sides, the total number of infection areas, and the approximate infected location. We followed the same division used by them. For QaTa-COV19, the train sets are split into 80% for training (5716 images) and 20% for validation (1429 images). For MosMedData, the split was about 8:1:1, resulting in 2183 training images, 273 validation images, and 273 test images.

## 3.2 Implementation Details

All methods were implemented by PyTorch. Additionally, to facilitate the implementation of baseline methods, we used PyTorch-Lightning to encapsulate the training and inference process. The MONAI library [22] was employed for implementing the up-sampling and the segmentation head in the Decoder Module. In addition, we used the pre-trained model ConvNeXt-tiny and CXR-BERT from Hugging Face as the Image and Text encoders, providing a high-performance backbone. The loss function $L$ is calculated by the sum of dice loss ($L_{Dice}$) and cross-entropy loss ($L_{Ce}$): $L = L_{Dice} + L_{Ce}$. The batch size, initial learning rate, and minimum learning rate were set to 32, 1e-5, and 1e-6, respectively. We chose the AdamW optimizer for network optimization and employed the cosine annealing learning rate policy for learning rate updates. All experiments were run on one Nvidia GeForce RTX 3090 with 24GB VRAM. Evaluation metrics for the model's segmentation

**Table 1.** Performance comparison of other state-of-the-art medical segmentation models on MosMedData+ and QaTa-COV19 test set

| Method | | MosMedData+ | | QaTa-COV19 | |
|---|---|---|---|---|---|
| | | Dice | MIoU | Dice | MIoU |
| W/O Text | U-Net [3] | 0.6460 | 0.5073 | 0.7902 | 0.6946 |
| | U-Net++ [4] | 0.7175 | 0.5839 | 0.7962 | 0.7025 |
| | nnUNet [5] | 0.7259 | 0.6036 | 0.8042 | 0.7081 |
| Text-Guided | C2FVL [17] | 0.7456 | 0.6115 | 0.8340 | 0.7462 |
| | CPAMTG [15] | - | - | 0.8425 | 0.7598 |
| | LViT [14] | 0.7457 | 0.6133 | 0.8366 | 0.7511 |
| | LGMS [16] | - | - | 0.8977 | 0.8145 |
| | TGCAM | **0.7782** | **0.6369** | **0.9060** | **0.8281** |

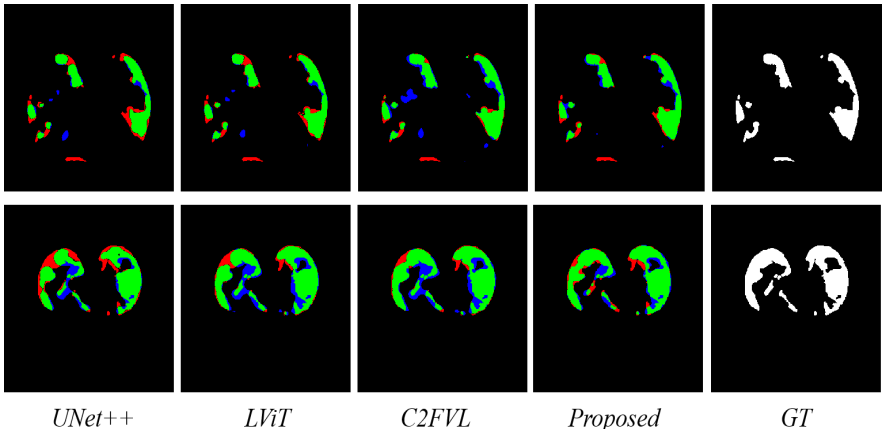


**Fig. 3.** Qualitative results of segmentation models on MosMedData+. Green, red and blue indicate true positive, false negative, and false positive pixels, respectively.

results include the Dice coefficient and the Mean Intersection over Union (MIoU) metric.

### 3.3 Comparison Experiments

Table 1 presents the comparison results of other state-of-the-art medical segmentation models on QaTa-COV19 and MosMedData+. We list common mono-modal segmentation methods UNet [3], UNet++ [4], nnUNet [5] (results aligned with LViT[14]) meanwhile almost all multi-modal methods for medical segmentation are chosen. By contrast, all methods without text show an apparent performance gap with text-guided multimodal methods, where nnUNet achieves the best performance of the mono-modal method, still 2.98% behind C2FVL [17] of Dice score, which demonstrates the ability of textual

**Table 2**. Ablation studies on the QaTa-COV19 test set. 'w/o text' means without text and the model use UNet Decoders only. ITEM, CAM represents Iterative Text Enhancement Module and Common Attention Module respectively

| No. | Model | QaTa-COV19 | |
|---|---|---|---|
| | | Dice | MIoU |
| #0 | Baseline (w/o text) | 0.8418 | 0.7262 |
| #1 | Baseline + ITEM | 0.8908 | 0.8062 |
| #2 | Baseline + CAM | 0.9022 | 0.8219 |
| #3 | Ours (Baseline + ITEM + CAM) | 0.9060 | 0.8281 |

information to guide segmentation. As for multi-modal methods, compared with C2FVL, our method improves Dice score by 7.2%, MIoU score by 8.19% on QaTa-COV19, which also surpasses CPAMTG [15] and LViT by a wide margin. LGMS [16] can reach quite good performance because of a similar baseline and the proposed method increases the dice score by just 0.83%, but we can achieve a 1.36% improvement in the MIoU score. On the MosMedData+, good performance is obtained, nearly 2.93% and 1.93% higher than LViT respectively for the Dice score and MIoU score.

The results of the qualitative experiment and ground truths are shown in Fig. 3. Green, red and blue, indicate true positive, false negative, and false positive pixels, respectively. The mono-modal methods segment many false negative pixels because of apparent performance gaps. Compared to other multi-modal methods, we reduce overmuch segment showed in blue apparently, while our proposed method maintains the same or even better ability to recognize true positive pixels.

### 3.4 Ablation Study

To validate the effectiveness of our proposed module. Ablation studies are conducted on the QaTa-COV19 data set shown in Table 2. The number represents different settings of the network. In Model #0, the baseline represents the image-only used method, whose results are treated as the basic reference. The Model#1 and Model#2 introduce the ITEM and CAM of our proposed method individually. They respectively bring a performance gain of 4.9% and 6.04%, which verifies the effectiveness of both modules. Finally, Model#3 which contains both modules achieves the Dice score of 90.6% and MIoU score of 82.81% as the best results.

## 4 Conclusion

We developed a text-guided segmentation method Text-Guided Common Attention Model (TGCAM), including multimodal information from text for automatic segmentation of infected regions of the lungs. We rethink the feature fusion pipeline depending on

attentional mechanisms and propose a genuine multimodal fusion attention mechanism named common attention. The designed Common Attention Block fuses text and image features at the decoding stage, where the Common Attention Module is responsible for the initial fusion and the Iterative Text Enhancement Module utilizes progressive conversions of text information for deeper multimodal interactions. The experimental results on both the X-ray and CT datasets indicate the advantages of our architecture compared to those image-only methods and multi-modal attention methods.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China (NSFC 62371325, 62071314), Sichuan Science and Technology Program 2023YFG0025, 2023YFG0101, and 2023 Science and Technology Project of Sichuan Health Commission 23LCYJ002.

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

## References

1. Torres, A., Cillóniz, C., Niederman, M.S., Menéndez, R., Chalmers, J.D., Wunderink, R.G., van der Poll, T.: Pneumonia. Nature Reviews Disease Primers 7, 1 (2021)
2. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., Bernardini, S.: The COVID-19 pandemic. Critical reviews in clinical laboratory sciences 57(6), 365-388 (2020)
3. Ronneberger O, Fischer P, Brox T, et al.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) MICCAI 2015, Part III 18, pp. 234-241. Springer, Cham (2015)
4. 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)
5. 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. Nature methods 18(2), 203-211 (2021)
6. Siddique, N., Paheding, S., Elkin, C. P., Devabhaktuni, V.: U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9, 82031-82057 (2021)
7. Kaiping Wang., Bo Zhan., Chen Zu., Xi Wu., Jiliu Zhou., Luping Zhou., Yan Wang.: Semi-supervised Medical Image Segmentation via a Tripled-uncertainty Guided Mean Teacher Model with Contrastive Learning. Medical Image Analysis, 79: 102447(2022)
8. Cheng Tang., Xinyi Zeng., Luping Zhou., Qizheng Zhou., Peng Wang., Xi Wu., Hongping Ren., Jiliu Zhou., Yan Wang.: Semi-supervised medical image segmentation via hard positives oriented contrastive learning. Pattern Recognition, 146: 110020(2024)
9. P Tang., P Yang., D Nie., X Wu., J Zhou., Y Wang.: Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowledge-Based Systems 241, 108215(2022)

10. Zeng X., Zeng P., Tang C., et al.: DBTrans: A Dual-Branch Vision Transformer for Multi-Modal Brain Tumor Segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 502-512. Springer, Cham (2023)
11. Radford A, Kim J W, Hallacy C, et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748-8763. PMLR, (2021)
12. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
13. Huang, S. C., Shen, L., Lungren, M. P., Yeung, S. 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, pp. 3942-3951. (2021)
14. Li Z, Li Y, Li Q, et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging, vol. 43, no. 1, 96-107 (2023)
15. Lee G E, Kim S H, Cho J, et al.: Text-Guided Cross-Position Attention for Segmentation: Case of Medical Image. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 537-546. Springer, Cham (2023)
16. Zhong Y, Xu M, Liang K, et al.: Ariadne's Thread: Using Text Prompts to Improve Segmentation of Infected Areas from Chest X-ray Images. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 724-733. Springer, Cham (2023)
17. Shan, D., Li, Z., Chen, W., Li, Q., Tian, J., Hong, Q.: Coarse-to-Fine Covid-19 Segmentation via Vision-Language Alignment. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE (2023)
18. Degerli, A., Kiranyaz, S., Chowdhury, M. E., Gabbouj, M.: Osegnet: Operational segmentation network for COVID-19 detection using chest X-ray images. In: IEEE International Conference on Image Processing (ICIP), pp. 2306-2310. IEEE (2022)
19. Morozov S P, Andreychenko A E, Pavlov N A, et al.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465. (2020)
20. 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)
21. Boecking B, Usuyama N, Bannur S, et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1-21. Springer, Cham (2022)
22. Cardoso M J, Li W, Brown R, et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701. (2022)