

Deco-CLIP: Dual-Branch Feature Decoupling CLIP for Fine-Grained Zero-Shot Medical Image Analysis

Tingjia Liu

Shanghai University of Electric Power

Ye Xia

Shanghai University of Electric Power

Xinran Yang

Shanghai University of Finance and Economics

Yusheng Liu

lys_sjtu@sjtu.edu.cn

Shanghai Jiao Tong University

Article

Keywords: Feature Decoupling, Cross-Modal Alignment, Zero-Shot

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10760166/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Deco-CLIP: Dual-Branch Feature Decoupling CLIP for Fine-Grained Zero-Shot Medical Image Analysis

Tingjia Liu¹, Ye Xia¹, Xinran Yang², Yusheng Liu^{3*}

¹School of Artificial Intelligence, Shanghai University of Electric Power,
Shanghai, 200090, China.

²School of Accountancy, Shanghai University of Finance and
Economics, Shanghai, 200433, China.

³Department of Automation, Shanghai Jiao Tong University, Shanghai,
200240, China.

*Corresponding author(s). E-mail(s): lys_sjtu@sjtu.edu.cn;
Contributing authors: 3560954412@qq.com; XiaYeWang@163.com;
1945697410@qq.com;

Abstract

The CLIP model has been widely used in the medical field, improving the efficiency of disease screening and diagnosis by establishing feature-level associations between medical images and radiology reports. However, we have found that the contrastive learning strategy of current medical CLIP primarily focuses on distinguishing high-level semantic features. Therefore, it fails to capture the low-level features related to attributes and relational information, limiting performance in fine-grained tasks. To address these limitations, we propose Deco-CLIP to decouple representational features from semantic features. Specifically, the image modality consists of two Vision Transformer (ViT) encoders that extract semantic and texture features using different pretraining strategies. The text modality consists of a text encoder with dual branches to focus on tokens representing high- and low-level features. Furthermore, to enhance the model’s ability to capture fine-grained features, we propose a threshold-adaptive sample mining strategy based on intra-modal similarity, which clusters semantically similar samples into mini-batches. Then, secondary contrastive learning is applied within the mini-batch to align the fine-grained features between image and text modalities. Additionally, a regularized attention mechanism is proposed to guide the two

that differentiate them from normal tissue. When non-lesional liver images are mixed with lesional liver image-text pairs for matching, the model fails to identify the lesional liver images. In Fig. 1(c), the correct spatial relationship between the liver and the hypodense lesion is "The hypodense lesion is observed at the liver." However, when the order of the liver and lesion is reversed in the text, the model is unable to accurately match text with image. In Fig. 1(d), the nodule in the image is located in the lower lobe of the left lung. When the attribute terms for "lobe" and "lung" are swapped in the text, the model similarly fails to correctly match the image to the text. In summary, as illustrated in Fig. 1(b), (c), and (d), the medical CLIP model usually exhibits limitations in low-level feature extraction, particularly in capturing image textures and variations in the order and attributes of text entities. These limitations hinder the model's performance in fine-grained medical tasks.

To address these limitations, we introduce Deco-CLIP, which decouples representational features from semantic features. For image modality, we use the Vision Transformer (ViT) encoders from pre-trained CLIP [16] and Masked Autoencoders (MAE)[23] to extract semantic and texture features, respectively. For text modality, we use a dual-branch text encoder, where each branch corresponds to one of the image encoders and is initialized randomly. To improve the model's ability to capture fine-grained features, we propose a threshold-adaptive sample mining strategy to separately cluster semantically similar image and text samples within a batch. In the clustered mini-batch, secondary contrastive learning is conducted by the MAE-trained image encoder and its corresponding text encoder branch. Additionally, to ensure the text encoder branches focus on distinct semantic tokens, a regularized attention mechanism is designed to refine the distribution of attention maps between the branches. We evaluated Deco-CLIP on fine-grained zero-shot tasks, including multi-abnormality detection and report-to-volume retrieval. Experimental results showed a 10.2% improvement in accuracy for multi-abnormality detection and a 16.2% increase in recall for report-to-volume retrieval, highlighting the effectiveness of Deco-CLIP in fine-grained medical tasks.

2 Method

In this section, we first introduce the architecture of Deco-CLIP. Then, we present the threshold-adaptive sample mining strategy and regularized attention mechanism. The details are provided in the following subsections.

2.1 The Architecture of Deco-CLIP

2.1.1 Image Encoder.

The image modality consists of two image encoders, E_a and E_b , as shown in Fig. 2(a). Of these, the E_a denotes the ViT from CT-CLIP[16], which is also pre-trained by CT-CLIP for the ability of high-level feature extraction. Additionally, the E_b denotes the ViT-B/16 and is pre-trained by the weights from MAE [23], as MAE is capable of extracting texture details via self-supervised learning and local masking strategy.

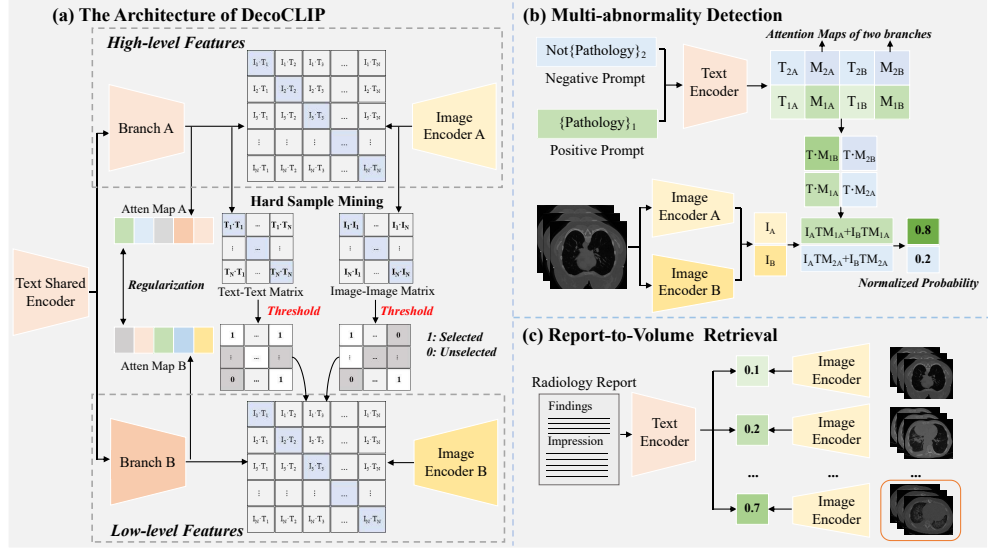


Fig. 2 (a) The architecture of Deco-CLIP consists of two pre-trained ViT encoders and a dual-branch text encoder. (b) Illustration of multi-abnormality detection using Deco-CLIP. (c) Illustration of report-to-volume retrieval using Deco-CLIP.

2.1.2 Text Encoder.

The text encoder is based on a 12-layer transformer architecture, modified with a dual-branch structure in the final three layers, denoted as B_a and B_b . Both branches share the same key (k) and value (v) vector, while using two distinct queries to focus on low- and high-level features, namely q_a and q_b . Based on the differences between q_a and q_b , two attention maps with different distributions are obtained by Eq. 1, denoted as A_a and A_b .

$$A_{\{a,b\}} = \sigma \left(\frac{q_{\{a,b\}} k^T}{\sqrt{d_k}} \right) v \quad (1)$$

where σ is the softmax function. Subsequently, the hidden layer features of B_a and B_b are obtained by the Transformer blocks $T(\cdot)$, denoted as H_a and H_b . Then, the outputs H_a^{final} and H_b^{final} are obtained by three consecutive Transformer blocks in their respective branches, as shown in Eq 2.

$$H_{\{a,b\}}^{final} = T_{\{a,b\}}^{\{i+2,j+2\}} \left(T_{\{a,b\}}^{\{i+1,j+1\}} \left(T_{\{a,b\}}^{\{i,j\}} (H_{\{a,b\}}, k, v, q_{\{a,b\}}) \right) \right) \quad (2)$$

where i and j are both equal to 10, indicating that the dual-branch structure begins from the 10-th layer, $T_{\{a,b\}}^{\{i,j\}}(\cdot)$ corresponds to the i -th and j -th transformer layer of B_a and B_b , and $H_{\{a,b\}}^{final}$ indicates the final output of both branches. Additionally, the text encoder is randomly initialized to prevent the pre-trained weights from influencing its ability to decouple high- and low-level features.

2.2 Contrastive Learning

2.2.1 Contrastive Learning for High-level Semantic Features.

To align high-level semantic features between image-text pairs, contrastive learning is used in Deco-CLIP to learn cross-modal representations from both positive and negative sample pairs. Specifically, the image features F_i^a and text features F_t^a are extracted by the image encoder E_a and the text encoder B_a . The features F_i^a and F_t^a are then aligned by calculating cosine similarity, as described in Eq 3.

$$\mathcal{L}_c = \frac{1}{|V_i^+|} \sum_{v^+ \in V_i^+} -\log \frac{e^{\text{sim}(v^+, v_i)/\tau}}{e^{\text{sim}(v^+, v_i)/\tau} + \sum_{v^- \in V_i^-} e^{\text{sim}(v^-, v_i)/\tau}} \quad (3)$$

where v_i represents the feature of the i -th sample, $v^+ \in V_i^+$ and $v^- \in V_i^-$ correspond to the features of its positive and negative samples. V_i^+ and V_i^- are the sets of features associated with the positive and negative samples of v_i . The $\text{sim}(a, b)$ represents the cosine similarity function and τ is the temperature parameter to control the concentration level of the distribution [24].

2.2.2 Threshold-Adaptive Sample Mining Strategy.

Building on the contrastive learning outlined in the previous section, which extracts key features from both images and text for alignment, we propose a sample mining strategy. This strategy adaptively constructs a mini-batch of semantically similar coarse-grained samples to further align fine-grained semantic features. Specifically, the intra-modal similarity matrices L_{ii} and L_{tt} are first computed from F_i^a and F_t^a using $\text{sim}(a, b)$. Subsequently, we define a threshold function $H(x; \theta)$, which equals 1 when x is greater than or equal to θ , and equals 0 when x is less than θ . Then, $H(x; \theta)$ is applied to L_{ii} and L_{tt} , resulting in binary mask matrices M_{ii} and M_{tt} . These masks are used to traverse each image or text sample and identify other samples within the same batch that exhibit high semantic similarity in the same modality. In this work, the θ is set to 0.75. As a result, M_{ii} and M_{tt} can be used to extract samples with high semantic similarity in the same modality.

2.2.3 Secondary Contrastive Learning for Low-level Representational Features.

To further improve the extraction of fine-grained features, secondary contrastive learning is applied to the MAE-trained visual encoder E_b and its corresponding text branch B_b . Specifically, the image features F_i^b and text features F_t^b , extracted by E_b and B_b , are used to compute image-text similarity via $\text{sim}(a, b)$. Unlike the loss $\mathcal{L}_c$ used for semantic features extraction, the secondary contrastive learning loss $\mathcal{L}_s$ computes the element-wise multiplication of the similarity matrix with the M_{ii} and M_{tt} matrices, thereby mining the image and text samples. Specifically, for each image sample, M_{ii} is used to mine other images with high intra-modal similarity. Cross-modal contrastive learning is then performed on these samples to extract fine-grained image features and align them with the corresponding text features, as demonstrated in Eq 4.

$$\mathcal{L}_{it} = \frac{-1}{\sum_{i=1}^N I\left(\sum_{j \neq i}^N M_{ii}[i, j]\right)} \sum_{i=1}^N \sum_{j \neq i}^N M_{ii}[i, j] \cdot \log \left(\frac{e^{\text{sim}(F_i^a, F_t^a[i])}}{\sum_{k=1}^N e^{\text{sim}(F_i^a, F_t^a[k])}} \right) \quad (4)$$

where N denotes the number of samples in the batch and $I(x)$ represents the indicator function, which equals 1 when x is greater than 0, and equals 0 when x is less than or equal to 0. Simultaneously, text samples with high semantic similarity are mined using M_{tt} , and the model’s ability to capture fine-grained text features is improved by contrastive learning on these samples, as shown in Eq 5.

$$\mathcal{L}_{ti} = \frac{-1}{\sum_{i=1}^N I\left(\sum_{j \neq i}^N M_{tt}[i, j]\right)} \sum_{i=1}^N \sum_{j \neq i}^N M_{tt}[i, j] \cdot \log \left(\frac{e^{\text{sim}(F_t^a, F_i^a[i])}}{\sum_{k=1}^N e^{\text{sim}(F_t^a, F_i^a[k])}} \right) \quad (5)$$

Finally, the secondary contrastive learning is optimized by $\mathcal{L}_s$, which averages $\mathcal{L}_{it}$ and $\mathcal{L}_{ti}$ to improve the model’s ability to extract and align fine-grained features across image and text modalities.

2.3 Loss Function

To ensure that the branches B_a and B_b focus on different semantic tokens, a regularized attention mechanism is proposed to guide the distributions of attention maps A_a and A_b , as shown in Eq 6. Specifically, $\text{sim}(a, b)$ is used to constrain the distributions of A_a and A_b to be orthogonal, ensuring that text features with different dimensions are extracted by branches B_a and B_b , respectively.

$$L_r = \frac{1}{N} \sum_{i=1}^N \text{sim} \left(\frac{a_{ai}}{\|a_{ai}\|}, \frac{a_{bi}}{\|a_{bi}\|} \right) \quad (6)$$

where N represents the batch size, $a_{ai} \in A_a$ and $a_{bi} \in A_b$ are the attention maps of the i -th text, $\|a_{ai}\|$ and $\|a_{bi}\|$ are the L2 norms of a_{ai} and a_{bi} .

Finally, the total loss for training Deco-CLIP is defined in Eq 7. After comprehensive hyperparameter tuning, we empirically set $\alpha = 0.4$, $\beta = 0.4$, and $\gamma = 0.2$.

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_c + \beta \cdot \mathcal{L}_s + \gamma \cdot \mathcal{L}_r \quad (7)$$

3 Experiences

3.1 Experimental Setup

We conducted extensive experiments on the CT-RATE dataset [16], which contains 3D chest CT volumes and corresponding radiology reports from 21,304 patients. Of these, 20,000 patients were used for the training set and 1,304 patients for the test set. Additionally, each CT volume is annotated with 18 types of abnormalities.

Table 1 Performance comparison between the proposed Deco-CLIP and other medical CLIP on multi-abnormality detection and report-to-volume retrieval tasks.

Method	Multi-Abnormality Detection			Report-to-Volume Retrieval			
	AUROC	Accuracy	Precision	R@5	R@10	R@50	R@100
CT-CLIP	0.702	0.651	0.304	0.023	0.040	0.144	0.235
CT-VocabFine	0.747	0.692	0.339	0.002	0.006	0.024	0.046
CT-LiPro	0.749	0.677	0.331	-	-	-	-
Deco-CLIP	0.782	0.753	0.423	0.025	0.045	0.192	0.397

To initialize Deco-CLIP, we used pre-trained weights from the encoders of CT-CLIP [16] and MAE [23] to initialize the image encoders E_a and E_b . In contrast, the text encoder was initialized randomly. To fine-tune Deco-CLIP, we randomly selected 30% of the samples from the training set, i.e., 6000 patients. The batch size was set to 10 and the learning rate to $1.25e-6$. Our fine-tuning strategy consists of two stages. In the first stage, we froze the parameters of E_a and E_b to train the text encoder for 210 epochs until convergence, achieving a loss value L_{total} of 0.31. In the second stage, we unfroze the last three layers of both E_a and E_b to further train the model for 600 epochs, achieving a L_{total} value of 0.1.

To evaluate the model’s ability to cross-modal alignment in fine-grained information, Deco-CLIP was applied to zero-shot tasks, including multi-abnormality detection and report-to-volume retrieval. For the multi-abnormality detection task, the probability of abnormalities is computed using positive (e.g., "Consolidation") and negative (e.g., "Not consolidation") prompts, as shown in Eq 8 and Fig. 2(b). These probabilities represent the likelihood of each abnormality’s presence or absence in a CT volume. For the report-to-volume retrieval task, we also calculate the probabilities between each report and all CT volumes by Eq 8. These probabilities are then used to rank the CT volumes and retrieve the most relevant matches, as shown in Fig. 2(c).

$$P_i = \sigma(f_{t_a} \cdot m_{t_a} \cdot f_{i_a} + f_{t_b} \cdot m_{i_b} \cdot f_{i_b}) \quad (8)$$

where P_i represents the probability of the i -th abnormality or report and σ is the softmax function. The f_{t_a} and f_{t_b} are the text features of the i -th text from B_a and B_b , with m_{t_a} and m_{t_b} as their attention maps. The f_{i_a} and f_{i_b} are the image features extracted by E_a and E_b , respectively. Additionally, we assess multi-abnormality detection performance using AUROC, accuracy, and precision, while evaluating report-to-volume retrieval by recalling the top 5, 10, 50, and 100 results, denoted as R@5, R@10, R@50, and R@100.

3.2 Experimental Results

To validate the capability of Deco-CLIP in fine-grained zero-shot medical tasks, we conducted validation on multi-abnormality detection and report-to-volume retrieval tasks, as shown in Table 1. Specifically, Deco-CLIP outperforms previous medical CLIP models in all metrics, achieving a 10.2% improvement in accuracy for multi-abnormality detection and a 16.2% increase in recall for report-to-volume retrieval.

<i>Case 1</i>								
B_a	Calibration	of	mediastinal	major	vascular	structures	is	natural
B_b	Calibration	of	mediastinal	major	vascular	structures	is	natural
<i>Case 2</i>								
B_a	Non-specific hypodense nodule	in	the	superior	right	lobe	of	the liver
B_b	Non-specific hypodense nodule	in	the	superior	right	lobe	of	the liver
<i>Case 3</i>								
B_a	A 2 mm diameter nodule	is	observed	in	the	lower	lobe	superior segment of the left lung
B_b	A 2 mm diameter nodule	is	observed	in	the	lower	lobe	superior segment of the left lung

Fig. 3 The visualization of attention maps A_a and A_b of the global tokens from the text encoder branches B_a and B_b . Darker shades of green indicate higher attention weights, while lighter shades represent lower attention weights.

Table 2 Ablation study on zero-shot tasks with Deco-CLIP using different image-text encoders.

<i>Image-Text</i>		<i>Multi-Abnormality Detection</i>			<i>Report-to-Volume Retrieval</i>			
E_a-B_a	E_b-B_b	<i>AUROC</i>	<i>Accuracy</i>	<i>Precision</i>	<i>R@5</i>	<i>R@10</i>	<i>R@50</i>	<i>R@100</i>
✓	-	0.712	0.704	0.409	0.021	0.042	0.155	0.325
-	✓	0.612	0.582	0.239	0.003	0.005	0.014	0.034
✓	✓	0.782	0.753	0.423	0.025	0.045	0.192	0.397

Furthermore, to validate the Deco-CLIP in decoupling representational features from semantic features, we visualized the attention maps A_a and A_b of the global tokens from the text encoder branches B_a and B_b , as shown in Fig. 3. Specifically, B_a , which focuses on high-level semantic features, assigns higher weights to tokens representing the semantic core of the sentence (e.g., “vascular structures,” “Non-specific hypodense nodule,” and “A 2 mm diameter nodule”). Conversely, B_b , which focuses on low-level representational features, emphasizes tokens related to attributes and spatial relationships (e.g., “natural,” “superior right lobe,” and “of the left lung”).

3.3 Ablation Study

We conducted an ablation study to assess the impact of each component in Deco-CLIP. Specifically, we evaluate its performance on two zero-shot tasks using the image-text encoders (E_a-B_a), the (E_b-B_b), and their combination (E_a-B_a and E_b-B_b). As shown in Table 2, although E_b-B_b is capable of capturing fine-grained features using the MAE encoder, its performance on both tasks is notably poor. However, our proposed Deco-CLIP combines E_a-B_a and E_b-B_b , which decouple the high- and low-level features, achieving the best performance.

4 Conclusion

We introduce Deco-CLIP, which consists of two image encoders and a dual-branch text encoder, to enhance the extraction of fine-grained features. To this end, we design a threshold-adaptive sample mining strategy. Additionally, a regularized attention mechanism is proposed and applied to the text branches to constrain the output distributions. Experimental results on zero-shot tasks demonstrate the effectiveness of Deco-CLIP in fine-grained medical image analysis.

Declarations

The research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

- [1] Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In: European Conference on Computer Vision, pp. 55–72 (2024). Springer
- [2] Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 2607–2615 (2024)
- [3] Liu, A., Guo, Y., Yong, J.-h., Xu, F.: Multi-grained radiology report generation with sentence-level image-language contrastive learning. IEEE Transactions on Medical Imaging (2024)
- [4] Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164 (2023)
- [5] Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., Zhou, X.: Mediclip: Adapting clip for few-shot medical image anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 458–468 (2024). Springer
- [6] Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)
- [7] Guo, M., Yi, H., Qin, Z., Wang, H., Men, A., Lao, Q.: Multiple prompt fusion for zero-shot lesion detection using vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 283–292 (2023). Springer

- [8] Mishra, A., Mittal, R., Jestin, C., Tingos, K., Rajpurkar, P.: Improving zero-shot detection of low prevalence chest pathologies using domain pre-trained language models. arXiv preprint arXiv:2306.08000 (2023)
- [9] Pham, T.T., Brecheisen, J., Nguyen, A., Nguyen, H., Le, N.: I-ai: A controllable & interpretable ai system for decoding radiologists’ intense focus for accurate cxr diagnoses. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7850–7859 (2024)
- [10] Van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 726–736 (2023). Springer
- [11] Dack, E., Brigato, L., McMurray, M., Fontanellaz, M., Frauenfelder, T., Hoppe, H., Exadaktylos, A., Geiser, T., Funke-Chambour, M., Christe, A., *et al.*: An empirical analysis for zero-shot multi-label classification on covid-19 ct scans and uncurated reports. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2634–2643 (2023)
- [12] Javed, S., Mahmood, A., Ganapathi, I.I., Dharejo, F.A., Werghi, N., Bennamoun, M.: Cclip: zero-shot learning for histopathology with comprehensive vision-language alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11450–11459 (2024)
- [13] Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11137–11146 (2024)
- [14] Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 420–429 (2023). Springer
- [15] Yu, X., Wu, Z., Zhang, L., Zhang, J., Lyu, Y., Zhu, D.: Cp-clip: Core-periphery feature alignment clip for zero-shot medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 88–97 (2024). Springer
- [16] Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasedelen, M.F., Wittmann, B., Simsar, E., Simsar, M., *et al.*: A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024)
- [17] Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H.: Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In: International Conference on Medical

- Image Computing and Computer-Assisted Intervention, pp. 403–412 (2023). Springer
- [18] Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 525–536 (2023). Springer
 - [19] Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 690–700 (2022). Springer
 - [20] Wang, P., Zhang, H., Yuan, Y.: Mcpl: Multi-modal collaborative prompt learning for medical vision-language model. *IEEE Transactions on Medical Imaging* (2024)
 - [21] You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 101–111 (2023). Springer
 - [22] Yu, X., Zhang, L., Wu, Z., Zhu, D.: Core-periphery multi-modality feature alignment for zero-shot medical image analysis. *IEEE Transactions on Medical Imaging* (2024)
 - [23] Lang, D.M., Schwartz, E., Bercea, C.I., Giryes, R., Schnabel, J.A.: 3d masked autoencoders with application to anomaly detection in non-contrast enhanced breast mri. *arXiv preprint arXiv:2303.05861* (2023)
 - [24] Chen, H., Wang, R., Wang, X., Li, J., Fang, Q., Li, H., Bai, J., Peng, Q., Meng, D., Wang, L.: Unsupervised local discrimination for medical images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(12), 15912–15929 (2023)