PolarDETR: Enhancing Interpretability in Multi-modal Methods for Jawbone Lesion Detection in CBCT

Yuxuan Yang[1], Chen Zhong[1], Xinyue Zhang[2], Ruohan Ma[2], Gang Li[2], Yong Guo[1], and Jupeng Li[1*]

[1] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
lijupeng@bjtu.edu.cn

[2] Department of Oral and Maxillofacial Radiology, Peking University School and Hospital of Stomatology, Beijing 100081, China

Abstract. Rapidly advancing multi-modal learning shows great promise in medical image analysis, but challenges remain in the detection of jawbone lesions. Existing general-purpose models fail to capture the relationships between anatomical contexts and spatial locations in CBCT images, and the complexity of these models hinders interpretability. We propose PolarDETR, a novel framework combining anatomical priors and multi-modal alignment through: 1) Polar Text-Position Encoding (PTPE), which links text to spatial coordinates via polar mapping, 2) Anatomical Constraint Learning, ensuring lesion detection within anatomically plausible regions, and 3) Position Matching Optimization for spatial consistency. Evaluated on 180 clinical cases (6929 CBCT slices), our method achieves a state-of-the-art mAP of 93.66%, outperforming both single-modal (e.g., DETR at 89.35%) and multi-modal models (e.g., CORA at 91.52%). Additionally, PolarDETR excels in interpretability, with an ACS of 84.12% and PMS of 80.45%, demonstrating its potential to enhance both detection performance and clinical usability in real-world applications. Our code is available at https://github.com/Cxxxsky/PolarDETR.

Keywords: Medical Image Analysis · Multi-modal Learning · PolarDETR · Anatomical Constrain.

1 Introduction

Jawbone lesion detection is crucial for early diagnosis and treatment planning in dental and maxillofacial medicine [17]. Lesions such as periapical lesions or cysts often appear in various forms on CBCT images, making their identification difficult. Accurate identification and localization are essential to prevent complications and improve outcomes, which makes advanced computational methods vital to clinicians.

[*] Corresponding author. Email: lijupeng@bjtu.edu.cn

Recent advancements in multi-modal learning have enhanced medical image analysis, particularly for lesion detection. By integrating visual data from CBCT images with textual information from clinical reports, multi-modal methods outperform traditional single-modal approaches, improving feature representation and enabling clinically relevant interpretation. Key innovations include ConVIRT [24], which used contrastive learning for cross-modal alignment; GLoRIA [8] and MGCA [12], which employed spatial attention for feature correspondence; and MedCLIP [20], which incorporated anatomical constraints to improve consistency. Additionally, BioViL [1] redefined text encoding with hybrid models, and MRM [5] introduced a low-level feature attenuation mechanism, benefiting downstream tasks. These advancements emphasize the growing significance of multi-modal learning in medical imaging.

Despite these advances, challenges remain in applying traditional multi-modal fusion methods to specialized tasks like jawbone lesion detection. 1) Ineffectiveness of General-Purpose Models: Common fusion methods, such as feature concatenation or Cartesian coordinate mapping [4], fail to consider the jawbone's anatomical geometry. These models, often pre-trained on natural images, struggle to capture specific features like trabecular structures, bone density variations, and lesion orientation [3], leading to incomplete lesion localization. 2) Reduced Interpretability with Increasing Complexity: The complexity of modern multi-modal models will decrease interpretability. Radiologists report difficulties in understanding how deep models reach conclusions, with 38% of Artificial Intelligence (AI)-generated findings distrusted due to lack of transparency [10]. This "black-box" issue obscures critical diagnostic information, such as whether a lesion is caused by bone density changes or adjacent tooth morphology.

To address these challenges, we propose PolarDETR, a novel method for jawbone lesion detection that enhances both accuracy and interpretability. By embedding clinical text-derived location information into the detection model's query space, PolarDETR aligns anatomical knowledge with CBCT images in a polar coordinate framework, improving lesion localization while maintaining interpretability. We also introduce an interpretability-as-a-service feature that pairs anatomical heatmaps with textual data, enabling clinicians to understand the rationale behind AI diagnoses and increasing trust in the model's output.

2 Methodology

2.1 Overall Framework

An overview of the proposed framework is shown in Fig. 1. Our model is based on the DEtection TRansformer (DETR) architecture for end-to-end object detection [2]. It includes three key components for jawbone lesion detection: **P**olar **T**ext-**P**osition **E**ncoding (**PTPE**), Anatomical Constraint Learning, and Position Matching Optimization. PTPE maps clinical text to spatial coordinates in the CBCT image, aligning anatomical locations with geometric features. Anatomical priors guide the detection process to ensure lesions are within plau-

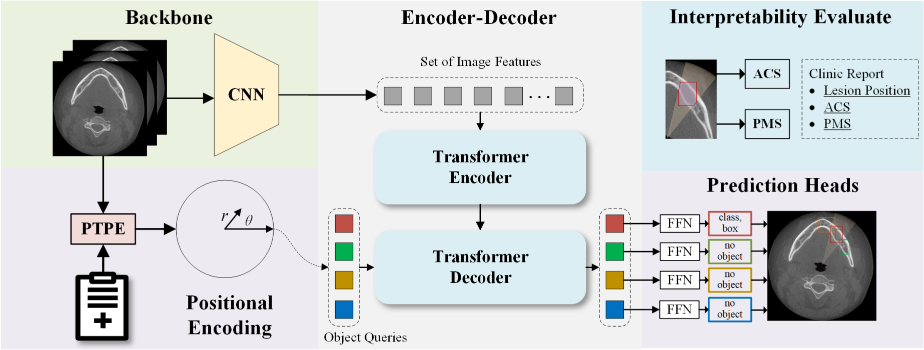

Fig. 1. The Overall Framework of Our PolarDETR.

sible regions, while Position Matching Optimization refines spatial alignment, improving accuracy and interpretability.

2.2 Polar Text-Position Encoding

To effectively incorporate anatomical location information from clinical text into image-based lesion detection in CBCT images, we introduce PTPE. This approach aligns the text and image representations in a shared feature space by leveraging the spatial geometry of anatomical regions described in the clinical text.

Anatomical Consistency with the Jaw Structure. The polar coordinate system demonstrates superior congruence with jaw anatomy compared to Cartesian frameworks. Oriented on radial anatomical patterns (e.g., orientation of the trabeculae, neurovascular bundles of the mental foramen), polar parameters (r, θ) inherently encode biomechanical characteristics lost in rectangular grids [13]. Clinically, this spatial representation directly maps to diagnostic descriptions (e.g., "5mm distal to mental foramen, third quadrant"), enhancing AI-output interpretability through native clinical lexicon alignment [14].

Design of PTPE. To construct PTPE, we first extract key positional information from the clinical text using a Named Entity Recognition (NER) model [11]. The PTPE pipeline includes an offline NER fine-tuning stage and an online inference stage integrated with PolarDETR. This model identifies anatomical terms (e.g., tooth numbers, quadrants) and their relationships. Critical elements like tooth number, quadrant, and distance from reference points (e.g., mental foramen), are parsed and converted into polar coordinates. The resulting PTPE is a 3D vector represented by θ (angle from the position of the main symptom) and r (radius from the description of the distance):

$$\text{PTPE} = [\sin(\theta), \cos(\theta), \log(r + 1)] \in \mathbb{R}^3 \tag{1}$$

To ensure geometrical consistency, a standard coordinate system with the chin foramen as the origin is defined, creating a polar coordinate system. To align image-based queries with anatomical information, PTPE vectors are integrated into DETR's object queries. This approach narrows the model's search space, focusing on anatomically relevant regions, accelerating convergence, and improving accuracy. The enhanced queries are computed as:

$$Q^{'} = Q + W_p \tag{2}$$

2.3 Anatomical Constraint and Position Matching Loss(AC-PML).

In this section, we introduce the AC-PML to improve the spatial accuracy of object localization and anatomical consistency in jawbone CBCT images. By combining anatomical understanding and precise alignment, we improve the localization of lesions and objects within the oral cavity.

Anatomical Constraint Learning. Anatomical constraint learning focuses on aligning image features with predefined anatomical regions. Most methods are handcrafted or use artificial prior information, lacking interpretability and ignoring the anatomical structure of complex regions, such as jawbone [19]. We define K anatomical regions, such as the alveolar bone and mandibular canal, using binary masks $\{M_k\}_{k=1}^{k}$ [22]. These masks serve as anatomical references for the model. For each image, we capture the relationship between image features and anatomical regions by normalizing the similarity between the query vector q and the average pooling of the anatomical mask M_k denoted as a_k. The association for each anatomical region is computed as follows:

$$\alpha_k = \frac{\exp(q^T a_k)}{\sum_{j=1}^{K} \exp(q^T a_k)}, a_k = \text{AvgPool}(M_k) \tag{3}$$

Anatomical Loss Function. To enforce anatomical consistency, we introduce an anatomical loss function based on Kullback-Leibler (KL) divergence [6] to minimize the distance between the predicted anatomical distribution α and the prior reflecting expected lesion locations α_{prior}. This approach uses anatomical priors to guide the model toward clinically plausible regions, ensuring detection focuses on anatomically relevant areas—crucial in jawbone CBCT images, where mislocalization will cause diagnostic errors. The anatomical loss is defined as:

$$L_{anatomy} = \text{KL}(\alpha || \alpha_{prior}) \tag{4}$$

Position Matching Learning. Previous research emphasizes the need to align text and visuals in tasks like lesion localization, as misalignment reduces performance [7]. Position-matching learning ensures spatial alignment between textual descriptions and corresponding image regions. Given the extracted polar coordinates (r, θ) from the text, we map these into a sector region R_{text} on the image.

The radial distance r is computed based on the pixel size and the text distance, while θ is the polar angle corresponding to the mapped tooth position. To align the predicted bounding box B_{pred} with the projected region R_{text}, we introduce a position-matching loss based on the Intersection over Union (IoU) [23] metric:

$$L_{position} = 1 - \mathrm{IoU}(B_{pred}, R_{text}) \tag{5}$$

The final loss function combines anatomical constraint learning and position matching learning, allowing more accurate and anatomically consistent predictions. By jointly optimizing these losses, the model is trained to focus on anatomical regions and spatial alignment, improving localization and anatomical understanding in jawbone CBCT image analysis.

2.4 Interpretability Metrics

In medical image analysis, integrating multi-modal clinical data requires careful evaluation of interpretability. To assess our lesion detection framework, we introduce two evaluation metrics: the **A**natomical **C**onsistency **S**core (**ACS**) and the **P**osition **M**atching **S**core (**PMS**). These metrics quantify the anatomical plausibility and spatial alignment between the clinical text and image regions. As shown in Fig. 2, after generating detection results, the model evaluates interpretability by comparing the detected lesions with both predefined anatomical regions and clinical semantic mappings.

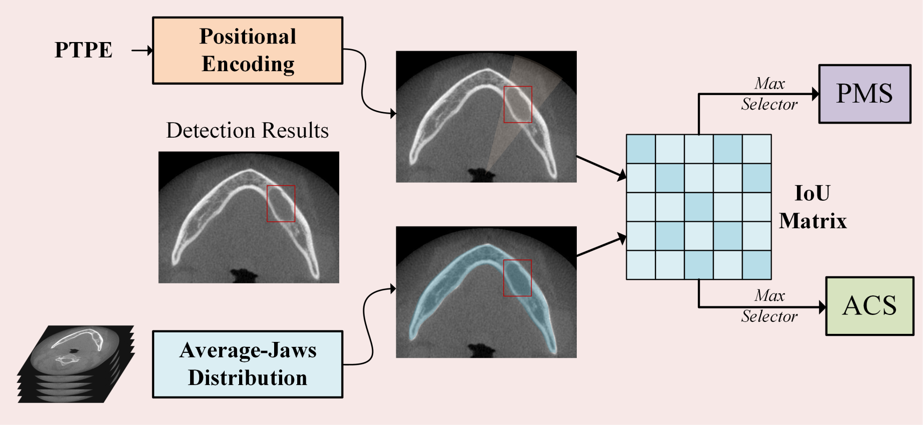

Fig. 2. Evaluation of Interpretability: Anatomical Consistency and Position Alignment.

Anatomical Consistency Score. The ACS quantifies the alignment between the predicted bounding boxes and anatomical regions. Specifically, ACS measures the IoU between each predicted bounding box and its corresponding anatomical region, taking the maximum IoU across all possible anatomical regions. The

ACS is computed as follows:

$$\text{ACS} = \frac{1}{N_{TP}} \sum_{N_{TP}}^{i=1} \max_k \text{IoU}(B_i, M_k) \tag{6}$$

where N_{TP} is the number of true positive predictions, B_i is the i-th predicted bounding box and M_k represents the k-th anatomical region corresponding to the i-th predicted box. $\text{IoU}(B_i, M_k)$ denotes the IoU between the predicted bounding box and the anatomical region. A higher ACS value indicates better anatomical plausibility of the model's predictions.

Position Match Score. The PMS measures the spatial consistency between the text descriptions and image predictions. This metric evaluates how well the predicted bounding boxes align with the anatomical regions described in the clinical text. The PMS is computed as:

$$\text{PMS} = \frac{1}{N_{TP}} \sum_{i=1}^{N_{TP}} \text{IoU}(B_i, R_{text,i}) > 0.5 \tag{7}$$

where B_i is the i-th predicted bounding box and $R_{text,i}$ is the reference region defined by the text description for the i-th sample. $\text{IoU}(B_i, R_{text,i})$ denotes the IoU between the predicted bounding box and the region specified by the text. A higher PMS value indicates better spatial alignment between the text and the detected regions.

By using both ACS and PMS, we provide a comprehensive evaluation of the anatomical consistency of the model and the spatial alignment between the clinical text and image data. These metrics contribute to the interpretability of our model, ensuring that it not only achieves high detection accuracy but also provides trustworthy and interpretable predictions that can aid clinicians in their decision-making process.

3 Experiments

3.1 Data Preparation and Preprocessing

In our experiments, we used a dataset of jawbone image-text pairs from Peking University School and Hospital of Stomatology, including CBCT slices (JPG format) and corresponding clinical texts. Each pair was linked to a unique patient. We excluded pairs with misleading information or poor quality (e.g., noise, artifacts). After quality control, the final dataset had 120 image-text pairs, each with a 512×512 pixel resolution, split into 70% for training 15% for validation, and 15% for testing. CBCT slices were standardized to 512×512 pixels, and bounding-box labels were created using LabelMe. For lesion detection, the ground truth (GT) includes the predicted bounding box (BBox), sector-shaped area (PMS), and average jaw distribution (ACS) based on the FDI tooth position method. For NER, GT includes tooth number, quadrant, and distance to anatomical reference points. To augment the clinical text data, we used a language model-based strategy to increase the number of image-text pairs, refined through expert collaboration.

3.2 Implementation Details

All experiments were conducted using the PyTorch 2.4 framework and executed on four Nvidia RTX 4090 GPUs. Our model was trained from scratch for a total of 200 epochs. A batch size of 4 was used on each GPU, with an initial learning rate set to 2e-4. We applied cosine decay, with the final learning rate reaching 2e-6 by the end of training.

3.3 Comparisons

We compared our model's performance with state-of-the-art image-only models, such as YOLOv8 [18], DETR [2], and Retina U-Net [9], as well as Image-Text models like CLIP [15] and CORA [21]. For a fair comparison, we used the default parameters from the open-source codes of competing methods, ensuring consistent data volume and training cycles. We evaluated accuracy and interpretability using mAP [16] (↑), ACS (↑), and PMS (↑).

The results in Table 1 show the detection and interpretability performance of single-modal vs. multi-modal models. Among single-modal models, DETR achieved the highest mAP of 89.35%, followed by YOLOv8 (87.26%) and Retina U-Net (85.11%). Multi-modal models outperformed single-modal ones, with CORA achieving a mAP of 91.52% and PolarDETR leading with the highest mAP of 93.66%. In terms of interpretability, PolarDETR outperformed others with an ACS of 84.12% and PMS of 80.45%, surpassing both single-modal and multi-modal models. Detection results for PolarDETR, shown in Fig. 3, highlight superior lesion localization and anatomical consistency, demonstrating the practical effectiveness and robustness of our approach in clinical applications.

Table 1. Performance comparison of different models in detection and interpretability.

Model	Model Type	mAP(%)	ACS(%)	PMS(%)
YOLOv8	Single-modal (Image-only)	87.26	72.30	N/A
DETR		89.35	75.48	N/A
Retina U-Net		85.11	70.87	N/A
CLIP	Multi-modal (Image-text)	84.37	78.56	73.21
CORA		91.52	81.43	74.58
PolarDETR		**93.66**	**84.12**	**80.45**

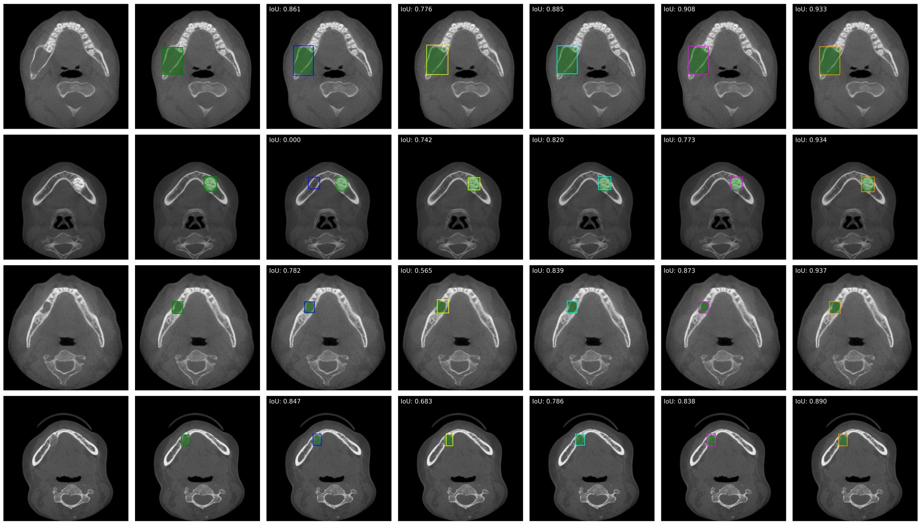

Fig. 3. Qualitative comparison of jawbone lesion detection results across different models. From left to right: original CBCT slices, GT, Retina U-Net, DETR, YOLOv8, CORA, and the proposed PolarDETR.

3.4 Ablation Study

We conducted ablation experiments to evaluate the effectiveness of the PTPE module and the AC-PML by training our model with and without these components. The first experiment compared the model with and without the PTPE module. The second experiment assessed the impact of AC-PML by comparing the full model with a version using only the standard DETR loss. Table 2 shows the results of these studies. Both the PTPE and AC-PML significantly improve detection accuracy, anatomical consistency, and position alignment.

Table 2. Ablation study results.

PTPE	AC-PML	mAP(%)	ACS(%)	PMS(%)
✗	✓	88.27	78.32	73.20
✓	✗	89.65	80.45	75.38
✓	✓	**93.66**	**84.12**	**80.45**

4 Conclusion

We proposed PolarDETR, a multi-modal approach that combines clinical text and CBCT images for jawbone lesion detection. With PTPE and AC-PML, our model enhances detection accuracy, anatomical consistency, and interpretability. Results show that PolarDETR outperforms existing models, offering greater interpretability for AI-assisted diagnosis. Future works will expand the dataset for better generalizability, explore integration with other imaging modalities, and focus on improving computational efficiency for clinical application.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62172029, the Beijing Natural Science Foundation under Grant L232029, and the Capital's Funds for Health Improvement and Research (CFH 2024-4-4107) from the Beijing Municipal Health Commission.

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

References

1. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229 (2020)
3. Chen, R.Q., Lee, Y., Yan, H., Mupparapu, M., Lure, F., Li, J., Setzer, F.C.: Leveraging pretrained transformers for efficient segmentation and lesion detection in cone-beam computed tomography scans. Journal of Endodontics **50**(10), 1505–1514 (2024)
4. Cho, H.: Cnn-based autoencoder and post-training quantization for on-device anomaly detection of cartesian coordinate robots. In: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference. pp. 0662–0666 (2024)
5. Colangelo, C.M., Chung, L., Bruce, C., Cheung, K.H.: Review of software tools for design and analysis of large scale mrm proteomic datasets. Methods **61**(3), 287–298 (2013)
6. Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., Zhang, H.: Decoupled kullback-leibler divergence loss. Advances in Neural Information Processing Systems **37**, 74461–74486 (2024)
7. Hossain, E., Sharif, O., Hoque, M.M., Preum, S.M.: Align before attend: aligning visual and textual features for multimodal hateful content detection. arXiv preprint arXiv:2402.09738 (2024)
8. 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)
9. Jaeger, P.F., Kohl, S.A., Bickelhaupt, S., Isensee, F., Kuder, T.A., Schlemmer, H.P., Maier-Hein, K.H.: Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In: Machine learning for health workshop. pp. 171–183 (2020)

10. Lastrucci, A., Wandael, Y., Ricci, R., Maccioni, G., Giansanti, D.: The integration of deep learning in radiotherapy: Exploring challenges, opportunities, and future directions through an umbrella review. Diagnostics **14**(9), 939 (2024)
11. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering **34**(1), 50–70 (2020)
12. Liu, M., Liu, Y., Cui, H., Li, C., Ma, J.: Mgct: Mutual-guided cross-modality transformer for survival outcome prediction using integrative histopathology-genomic features. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine. pp. 1306–1312 (2023)
13. Mei, S., Jiang, R., Ma, M., Song, C.: Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023)
14. Pelé, A., Berry, P.A., Evanno, C., Jordana, F.: Evaluation of mental foramen with cone beam computed tomography: a systematic review of literature. Radiology research and practice **2021**(1), 8897275 (2021)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
16. Shah, T.: Measuring object detection models—map—what is mean average precision. Tarang Shah-Blog **26**, 104332 (2018)
17. Shi, Y.J., Li, J.P., Wang, Y., Ma, R.H., Wang, Y.L., Guo, Y., Li, G.: Deep learning in the diagnosis for cystic lesions of the jaws: a review of recent progress. Dentomaxillofacial Radiology **53**(5), 271–280 (2024)
18. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems. pp. 1–6 (2024)
19. Wang, X., Li, X., Du, R., Zhong, Y., Lu, Y., Song, T.: Anatomical prior-based automatic segmentation for cardiac substructures from computed tomography images. Bioengineering **10**(11), 1267 (2023)
20. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
21. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7031–7040 (2023)
22. Xie, Y., Yang, B., Guan, Q., Zhang, J., Wu, Q., Xia, Y.: Attention mechanisms in medical image segmentation: A survey. arXiv preprint arXiv:2305.17937 (2023)
23. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 516–520 (2016)
24. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25 (2022)