



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# GEM: Context-Aware Gaze EstiMation with Visual Search Behavior Matching for Chest Radiograph

Shaonan Liu<sup>1</sup>, Wenting Chen<sup>2\*</sup>, Jie Liu<sup>2</sup>, Xiaoling Luo<sup>1,4</sup> and Linlin Shen<sup>1,3,4</sup> \*

<sup>1</sup>Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University

<sup>2</sup>Department of Electrical Engineering, City University of Hong Kong

<sup>3</sup>AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen University

<sup>4</sup>Guangdong Provincial Key Laboratory of Intelligent Information Processing

**Abstract.** Gaze estimation is pivotal in human scene comprehension tasks, particularly in medical diagnostic analysis. Eye-tracking technology facilitates the recording of physicians' ocular movements during image interpretation, thereby elucidating their visual attention patterns and information-processing strategies. In this paper, we initially define the context-aware gaze estimation problem in medical radiology report settings. To understand the attention allocation and cognitive behavior of radiologists during the medical image interpretation process, we propose a context-aware Gaze EstiMation (GEM) network that utilizes eye gaze data collected from radiologists to simulate their visual search behavior patterns throughout the image interpretation process. It consists of a context-awareness module, visual behavior graph construction, and visual behavior matching. Within the context-awareness module, we achieve intricate multimodal registration by establishing connections between medical reports and images. Subsequently, for a more accurate simulation of genuine visual search behavior patterns, we introduce a visual behavior graph structure, capturing such behavior through high-order relationships (edges) between gaze points (nodes). To maintain the authenticity of visual behavior, we devise a visual behavior-matching approach, adjusting the high-order relationships between them by matching the graph constructed from real and estimated gaze points. Extensive experiments on four publicly available datasets demonstrate the superiority of GEM over existing methods and its strong generalizability, which also provides a new direction for the effective utilization of diverse modalities in medical image interpretation and enhances the interpretability of models in the field of medical imaging. <https://github.com/Tiger-SN/GEM>

## 1 Introduction

Eye tracking is a key technology providing gaze to understand human behavior and fundamental cognitive processes, with widespread applications in different

---

\* Corresponding authors: Wenting Chen ([wentichen7-c@my.cityu.edu.hk](mailto:wentichen7-c@my.cityu.edu.hk)) and Linlin Shen ([llshen@szu.edu.cn](mailto:llshen@szu.edu.cn))

areas [25,11]. Several studies have explored gaze data in various radiology fields, encompassing examinations such as Chest CT [1], and Knee X-rays [27]. These studies provide insights into radiologists’ accurate attention allocation and cognitive behavior during image interpretation, and valuable information for understanding and diagnosing. For instance, the distribution and density of gaze data can reveal attention patterns, allowing for precise identification of each potential lesion with a single gaze point to inform diagnostic decisions. Moreover, eye tracking can assist surgeons in dynamically adjusting their perspectives during procedures [14]. The recordings of eye movements can also be replayed as training materials. Thus, monitoring the eye movement strategies of clinicians during image interpretation is crucial for lesion location and clinical disease diagnosis.

Automatic gaze estimation algorithms for natural scenes have been investigated in recent years. Some methods [7,15] typically leverage the scene image and head image to detect gaze regions, and others [26,23,22] utilize multi-task prediction (e.g. object detection and category classification) to enhance gaze estimation accuracy. Although these methods achieve significant gaze estimation performance in natural scenes, they cannot be directly applied to clinical practice. Because radiologists often encounter difficulty in obtaining head images during medical image interpretation, models are unable to utilize crucial information regarding head tracking. In contrast, the paired reports and images are readily available, and eye gaze data can easily establish their connections. Furthermore, with the success of language models, some works have integrated pre-trained language models into medical downstream tasks [17,29,5,6,28]. This inspires us to explore visual behavior patterns by establishing relationships between images and texts through eye-tracking data. Additionally, radiologists typically exhibit inherent behavioral patterns when interpreting images. For instance, they tend to first focus on prominent key lesion locations before scanning surrounding areas [12,4]. Therefore, we have raised two main issues for the medical gaze estimation method. **Q1**: Can we establish a connection between textual content and visual perception to achieve a more accurate localization of complex and various lesions in the medical image? **Q2**: Can we design a new model architecture to simulate the visual search behavior patterns of radiologists, and thereby gain deeper insights into the visual strategies and decision-making processes of radiologists during image interpretation?

To address the aforementioned issue, we propose a **Context-Aware Gaze EstiMation(GEM)** network for medical images, which first defines a medical-specific gaze estimation problem. GEM aims to predict the eye gaze of a radiologist with the given medical image and lesion name. **First**, to establish the connection between lesion names and medical images, we devise a **context-aware module** to achieve the fine-grained multi-modal alignment. With the fine-grained alignment, we can provide a precise textual clue for accurate gaze estimation. **Second**, to better simulate real visual search behavior patterns, we propose to use graph matching to capture and preserve the relationships among gaze points during gaze estimation. To be specific, we introduce the **visual behavior graph construction** to capture visual search behavior by representing

it with the high-order relations (i.e. edge) among gaze points (i.e. nodes). To preserve the real visual behavior, we design a **visual behavior matching** to align the high-order relation between the real and estimated gaze points by matching their constructed graphs. With these devised modules, the proposed GEM can precisely locate the specific position of the given lesion name, thus offering vital interpretation for radiologists’ auxiliary analysis. Experimental results on four publicly available Chest X-ray datasets show the superiority of GEM to existing methods and exhibit impressive generalizability across both easy and hard tasks with the zero-shot setting.

## 2 Method

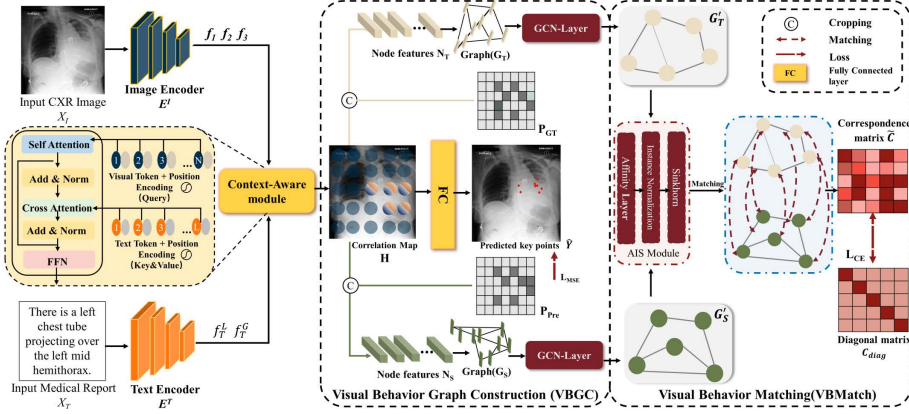
### 2.1 Problem Definition

Let  $\{(X_I^i, X_T^i, Y^i)\}_{i=1}^N$  denotes the dataset, where  $X_T$  represents the sentence of the medical report,  $X_I \in \mathbb{R}^{H \times W}$  represents chest X-ray (CXR) images.  $H$ , and  $W$  denote the height, and width of the CXR image, respectively.  $Y \in \mathbb{R}^{K \times 2}$  stands for the coordinates  $(x, y)$  of  $K$  gaze points for a given sentence. The objective of this task is to minimize the function:  $\arg \min_{\theta} \frac{1}{2N} \sum_{i=1}^N (F(X_I^i, X_T^i; \theta) - Y^i)^2$ .  $F$  represents the function for predicting gaze points  $Y$ , where  $\theta$  is its corresponding parameter.

### 2.2 Context-Aware Gaze Estimation

In Fig. 1, we propose a Context-Aware Gaze EstiMation (GEM) network to predict the eye gaze points for the medical reports. Given a CXR image  $X_I$  and a sentence  $X_T$  of the corresponding medical report, the image and text encoder  $E^I, E^T$  encode them to visual and textual features, respectively. Next, the context-aware module takes these features as input and generates a correlation map  $H$  to show the fine-grained multi-modal alignment relation. Then,  $H$  is passed to a fully connected layer to predict gaze points  $\hat{Y}$ . In the visual behavior graph construction, to represent the behavior, we prepare the gaze masks of the ground truth and predicted gaze points  $P_{GT}, P_{Pre}$ , and crop the correlation map with gaze masks to obtain node features  $N_T, N_S$  for their graphs  $G_T, G_S$ . To preserve visual behavior, we perform visual behavior matching by computing the correspondence matrix between  $G_T$  and  $G_S$  through AIS module and forcing it close to the diagonal matrix.  $L_{MSE}$  and  $L_{CE}$  are used for optimization. **Image Encoder.** We adopt ResNet-50 as the backbone of our image encoder to obtain three separate features  $f_1 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_1}, f_2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}, f_3 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_3}$ , where  $C_1, C_2, C_3$  denotes the dimension of each feature, and  $H$  and  $W$  denote the height and width of the CXR image, respectively.

**Text Encoder.** We utilize the text encoder pre-trained by CLIP [19] to extract the global textual features  $f_T^G \in \mathbb{R}^D$  for the input text and the local textual features  $f_T^L \in \mathbb{R}^{M \times D}$  for each textual token, where  $D$  and  $M$  denote the dimension of textual features and the number of textual tokens, respectively.



**Fig. 1. Overview of the proposed Context-Aware Gaze Estimation (GEM) network.** It consists of a context-aware module for fine-grained inter-modal alignment, a visual behavior graph construction to capture radiologists’ visual search behavior, and a visual behavior matching module to preserve the behavior.

**Context-Aware Module.** To capture the inter-modal relation in different scales, we fuse textual features with multi-scale visual features. Concretely, we first perform element-wise multiplication of  $Conv(f_1)$  and  $MLP(f_T^G)$  and then upscale by the factor of 2 to obtain the high-level integrated feature  $\mathcal{F}_1$ , where  $Conv$  and  $MLP$  denote the convolutional layer with kernel size of  $1 \times 1$  and the multilayer perceptron. For further fusion in middle scale, we integrate  $f_2$  with  $\mathcal{F}_1$  to obtain intermediate-level integrated features  $\mathcal{F}_2 = UP(Conv(f_2 \cdot \mathcal{F}_1))$ , where  $UP$  and  $\cdot$  denote  $\times 2$  upsampling layer and concatenation operation, respectively. Next, we obtain the low-level integrated features  $\mathcal{F}_3 = AVG(Conv(f_3 \cdot \mathcal{F}_2))$ , where  $AVG$  denote downsampling layer. With multi-level integrated features, we aggregate them through CoordConv layer [18] to generate the multi-modal features  $\mathcal{F}_m$ . Finally, we pass  $\mathcal{F}_m$  and local textual features  $f_T^L$  to self-attention layer  $SA$ , cross-attention layer  $CA$  and feed-forward network  $FFN$  to obtain the correlation map  $H = FFN(CA(SA(\mathcal{F}_m + PE, f_T^L + PE)))$ , where  $PE$  indicates the position embeddings. The correlation map shows the fine-grained alignment relation between the input CXR image and texts.

### 2.3 Visual Behavior Graph Construction (VBGC)

As radiologists’ gaze points scatter similar to a star pattern with a central point and surrounding ones, we introduce visual behavior graph construction to capture their visual behavior, which utilizes a graph to represent the high-order relations [16] (i.e., edges) among gaze points (i.e., nodes). Specifically, we prepare the gaze masks for the ground-truth (GT) and estimated gaze points  $P_{GT}, P_{Pre}$ . The gaze mask is obtained by masking  $6 \times 6$  patches centered around the gaze points. To extract the gaze point features, we crop the correlation map with the

gaze masks to obtain the node features for the GT and estimated gaze points  $N_T, N_S$ . With the node features, we obtain their graph edges  $E_T$  and  $E_S$  by applying an edge generator [10] to the node features. The edge generator initially employs a transformer model to learn the soft edge relationships between any pair of nodes in the graph. Then, the softmax function is applied to the inner product of soft edge features to derive the soft edge adjacency matrices  $E_T$  and  $E_S$ . These graph edges unveil the higher-order relationships between the GT and estimated gaze point features, facilitating a more profound comprehension of the interaction among features across various levels. After that, we obtain the graph structures  $G_T = \{N_T; E_T\}$  and  $G_S = \{N_S; E_S\}$  for GT and estimated gaze points, respectively. To capture higher-order relationships, we utilize graph convolutional networks (GCN) to embed graph nodes and higher-order graph structures (edges) into the node feature space, thereby generating new node features  $GCN(N_T, E_T)$  and  $GCN(N_S, E_S)$ .

#### 2.4 Visual Behavior Matching (VBMatch)

To learn the actual visual behavior of radiologists, we propose a visual behavior matching to align the high-order relation between the GT and estimated gaze points by matching their graphs. To be specific, given the graphs for GT and estimated gaze points  $G_T, G_S$ , we perform graph matching between them to reduce the disparity in their corresponding relationships within the graph feature space. The AIS module [10] is employed to calculate the soft correspondence matrix,

$$\tilde{C} = \text{AIS}(GCN(N_T, E_T), GCN(N_S, E_S)), \quad (1)$$

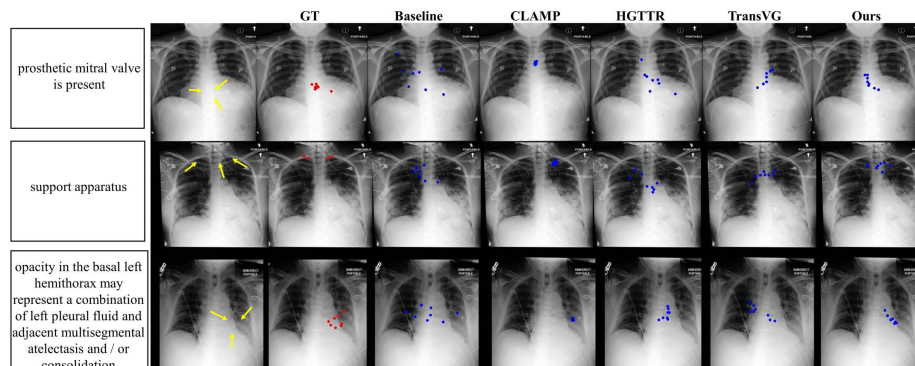
where  $\tilde{C}$  represents the soft correspondence between nodes in two predicted graphs, indicating the likelihood of establishing matching relationships between any pair of nodes in the two graphs. The AIS module [10] consists of an affinity layer to compute an affinity matrix between two graphs, instance normalization [24] to ensure the elements of the affinity matrix are positive, and Sinkhorn [20] to address outliers in the affinity matrix. To encourage the high-order relation of GT gaze points close to that of estimated ones, we force one-to-one correspondence between their nodes and edges by setting the target relation matrix as a diagonal matrix. To achieve this, we compute the cross-entropy loss  $L_{CE}$  between the soft correspondence matrix  $\tilde{C}$  and the target correspondence matrix  $C_{diag}$ . In order to provide more supervision to the estimated gaze points, we employ the Mean Squared Error (MSE) loss  $L_{MSE}$  to quantify the disparity between them. The overall objective function is defined as:

$$L_{\text{total}} = \alpha L_{MSE}(\hat{Y}, Y) + \beta L_{CE}(\tilde{C}, C_{diag}). \quad (2)$$

### 3 Experiments and Results


#### 3.1 Experiment Setting

**Datasets.** We conduct the experiments on four publicly available chest X-ray (CXR) datasets, including **MIMIC-Eye** [13], **OpenI** [8], **MS-CXR** [2] and



**Fig. 2.** Qualitative comparison of GEM and other models on MIMIC-Eye dataset. Yellow arrows indicate regions relevant to input texts, and red and blue points represent the GT and estimated gaze points, respectively.

**Table 1.** Gaze estimation performance of different methods. Asterisks indicate statistical significance: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

Methods	MSE	MAE	PCK@0.2	PCK@0.3	PCK@0.4
Baseline	0.0387***	0.156***	41.950***	68.701***	87.058***
CLAMP [31]	0.0580***	0.182***	37.549***	57.806***	74.035***
HGTR [23]	0.0357***	0.150***	43.751***	71.174***	89.528**
TransVG [9]	0.0327*	0.141*	50.032*	74.951**	89.217**
 GEM	<b>0.0320</b>	<b>0.139</b>	<b>50.490</b>	<b>77.012</b>	<b>90.477</b>

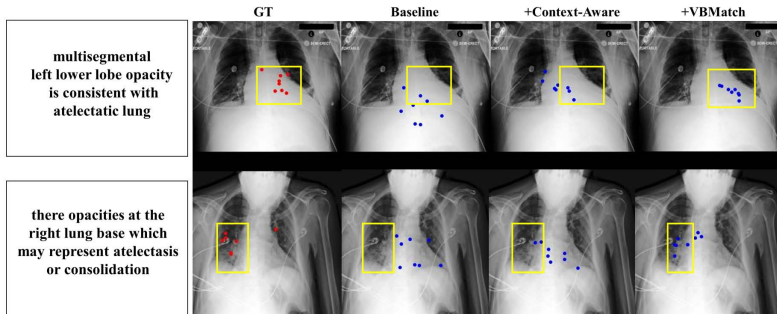
**AIforCovid** [21]. The **MIMIC-Eye** dataset includes 8,164 pairs of CXR images, diagnostic reports, and eye-tracking data collected from radiologists while examining diagnostic reports and images from the cohort of 3,192 patients. 6,600 pairs are used for training, 800 for validation, and 764 for testing. The **OpenI** dataset comprises 3,684 pairs of radiology reports and CXR images, while the **AIforCovid** dataset consists of the clinical data and CXR images from 820 registered patients. The **MS-CXR** dataset contains 1,047 CXR images with annotations of 1,153 bounding boxes and the corresponding sentences. The CXR images are resized to  $224 \times 224$ .

**Implementation Details.** We adopt the image and text encoder pre-trained by CLIP and freeze the parameters during training. AdamW optimizer is used with a learning rate of  $1e-6$  and a batch size of 16. The training epoch is 12.  $\alpha$  and  $\beta$  are set to 1 and 0.1. All experiments are conducted with the PyTorch on one Nvidia V100 32GB GPU.

**Evaluation Metrics.** We utilize Mean Squared Error (MSE), Mean Absolute Error (MAE) distances, and the Probability of Correct Keypoint (PCK) [30] to assess gaze estimation accuracy. PCK incorporates predefined thresholds of 0.2, 0.3, and 0.4 for a thorough and balanced evaluation.

**Table 2.** Ablation study on the MIMIC-Eye dataset.

Baseline	GEM		Metrics				
Addition	Context-Aware	VBMatch	MSE	MAE	PCK@0.2	PCK@0.3	PCK@0.4
✓	✗	✗	0.0387	0.156	41.950	68.701	87.058
✗	✓	✗	0.0322	0.142	48.020	75.867	87.009
✗	✓	✓	<b>0.0320</b>	<b>0.139</b>	<b>50.490</b>	<b>77.012</b>	<b>90.477</b>

**Fig. 3.** Visualization of the ablation study on the MIMIC-Eye dataset.

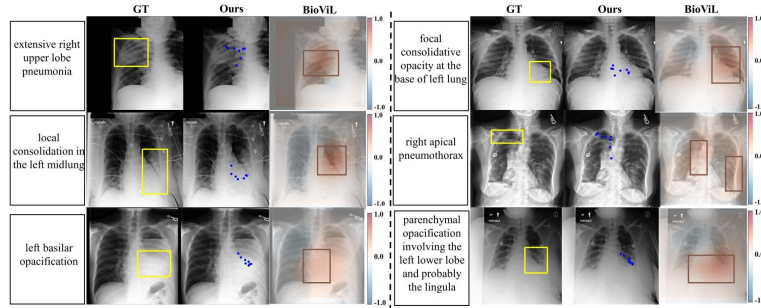
### 3.2 Comparison with Baseline Models

We quantitatively and qualitatively compare our performance with the baseline model and keypoints estimation methods for natural scenes, including the CLAMP [31] that employs text cues to estimate animal poses TransVG [9] whose output layer is modified for gaze estimation, and HGGTR [23] with text input. In Table 1, our GEM surpasses current methods by a large margin with the PCK@0.3 scores of 19.206 and 2.061, exhibiting the superiority of the proposed method. Fig. 2 visualizes the gaze points estimated by the proposed method and other methods. The baseline model produces scattered predictions, CLAMP aggregates gaze into a single position, but TransVG’s and HGGTR’s estimations lack accuracy. Conversely, GEM not only predicts precise gaze points but also captures radiologists’ visual search behavior, demonstrating our superiority.

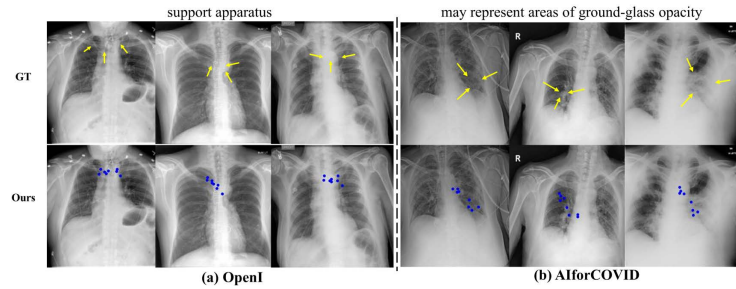
### 3.3 Ablation Study

**Effectiveness of Context-Aware Modules.** In Table 2, the baseline model integrates textual to visual features through element-wise addition for gaze estimation. When using context-aware module, the performance is significantly improved by 10.166 in the PCK@0.3 score. Fig. 3 shows that the gaze points (blue) predicted by the baseline with context-aware module are closer to the GT ones (red), indicating the effectiveness of the context-aware module.

**Effectiveness of Visual Behavior Matching (VBMatch).** After integrating both visual behavior matching (VBMatch) and context-aware module, the PCK@0.3 score remarkably increases by 1.145, as listed in Table 2. Fig. 3 illustrates that further equipping the baseline model with VBMatch makes the



**Fig. 4.** Qualitative evaluation of our method on the easy task of phrase grounding on the MS-CXR dataset. Yellow boxes indicate radiologists’ annotations, blue points are estimated points, and the heatmap and box show focal areas highlighted by BioViL [3].



**Fig. 5.** Qualitative evaluation of our method on the hard task of gaze estimation on OpenI and AIforCOVID datasets. Yellow arrows indicate radiologists’ annotation, and blue points represent estimated gaze points.

estimated points more accurate and similar to actual search behavior, suggesting the effectiveness of VBMatch.

### 3.4 Generalizability Analysis

To further analyze the generalizability of our method, we design both easy and hard tasks with the zero-shot setting. The former is the phrase grounding task that predicts the rough region (i.e. bounding box) with given texts on the MS-CXR dataset [2], and the latter is to estimate precise gaze points on the OpenI [8] and AIforCovid [21] datasets. In the easy task, Fig. 4 shows that the gaze points generated by our method are accurately located in the GT box and comparable to detection results of BioViL [3], indicating that our method can generalize well on the phrase grounding task. For the hard task, in Fig. 5, our method can generate precise gaze points that highly correspond to the real annotations on both unseen datasets, implying our strong generalizability in zero-shot settings.



## 4 Conclusion

In this work, we propose a GEM network, which presents the first work to define and address the context-aware gaze estimation problem in medical scenarios. We devise a context-aware module to establish the fine-grained relation between medical images and reports to facilitate gaze estimation. To learn the radiologists’ visual search behavior patterns, we propose visual behavior graph construction to represent the visual behavior with graphs and employ visual behavior matching to preserve the behavior. Extensive experiments prove the superiority of the proposed method over other models and show its strong generalizability and interpretability across easy and hard tasks with zero-shot settings.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China under Grant 82261138629 and 12326610; Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010688; Guangdong Provincial Key Laboratory under Grant 2023B1212060076; Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030.

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

## References

1. Aresta, G., Ferreira, C., Pedrosa, J., Araujo, T., Rebelo, J., Negro, E., Morgado, M., Alves, F., Cunha, A., Ramos, I., et al.: Automatic lung nodule detection combined with gaze information improves radiologists’ screening performance. *IEEE J. Biomed. Health Inform.* **24**(10), 2894–2901 (2020)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: *ECCV*. pp. 1–21. Springer (2022)
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: *ECCV*. pp. 1–21. Springer (2022)
4. Brunyé, T.T., Nallamothu, B.K., Elmore, J.G.: Eye-tracking for assessing medical image interpretation: A pilot feasibility study comparing novice vs expert cardiologists. *Perspectives on medical education* **8**, 65–73 (2019)
5. Chen, W., Li, X., Shen, L., Yuan, Y.: Fine-grained image-text alignment in medical imaging enables cyclic image-report generation. *arXiv preprint arXiv:2312.08078* (2023)
6. Chen, W., Wang, P., Ren, H., Sun, L., Li, Q., Yuan, Y., Li, X.: Medical image synthesis via fine-grained image-text alignment and anatomy-pathology prompting. *arXiv preprint arXiv:2403.06835* (2024)
7. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: *ECCV*. pp. 383–398 (2018)

8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* **23**(2), 304–310 (2016)
9. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: *ICCV*. pp. 1769–1779 (2021)
10. Fu, K., Liu, S., Luo, X., Wang, M.: Robust point cloud registration framework based on deep graph matching. In: *CVPR*. pp. 8893–8902 (2021)
11. Van der Gijp, A., Ravesloot, C., Jarodzka, H., Van der Schaaf, M., Van der Schaaf, I., van Schaik, J.P., Ten Cate, T.J.: How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Adv Health Sci Educ Theory Pract* **22**, 765–787 (2017)
12. Henderson, J.M.: Human gaze control during real-world scene perception. *Trends in cognitive sciences* **7**(11), 498–504 (2003)
13. Hsieh, C., Ouyang, C., Nascimento, J.C., Pereira, J., Jorge, J., Moreira, C.: Mimic-eye: Integrating mimic datasets with reflax and eye gaze for multimodal deep learning applications (2023)
14. Ikeda, A., Izumi, K., Katori, K., Nosato, H., Kobayashi, K., Suzuki, S., Kandori, S., Sanuki, M., Ochiai, Y., Nishiyama, H.: Objective evaluation of gaze location patterns using eye tracking during cystoscopy and artificial intelligence-assisted lesion detection. *J. Endourol.* (2024)
15. Lian, D., Yu, Z., Gao, S.: Believe it or not, we know what you are looking at! In: *ACCV*. pp. 35–50. Springer (2018)
16. Liu, J., Guo, X., Yuan, Y.: Graph-based surgical instrument adaptive segmentation via domain-common knowledge. *IEEE Trans. Med. Imag.* **41**(3), 715–726 (2021)
17. 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: *ICCV*. pp. 21152–21164 (2023)
18. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. *NeurIPS*. **31** (2018)
19. 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: *ICML*. pp. 8748–8763 (2021)
20. Sinkhorn, R.: A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics* **35**(2), 876–879 (1964)
21. Soda, P., D’Amico, N.C., Tessadori, J., Valbusa, G., Guarrasi, V., Bortolotto, C., Akbar, M.U., Sicilia, R., Cordelli, E., Fazzini, D., et al.: Aiforcovid: Predicting the clinical outcomes in patients with covid-19 applying ai to chest-x-rays. an italian multicentre study. *Med. Image Anal.* **74**, 102216 (2021)
22. Tonini, F., Dall’Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: *ICCV*. pp. 21860–21869 (2023)
23. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: *CVPR*. pp. 2192–2200. IEEE (2022)
24. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016)
25. Waite, S., Grigorian, A., Alexander, R.G., Macknik, S.L., Carrasco, M., Heeger, D.J., Martinez-Conde, S.: Analysis of perceptual expertise in radiology—current knowledge and a new perspective. *Front. Hum. Neurosci.* **13**, 213 (2019)
26. Wang, B., Hu, T., Li, B., Chen, X., Zhang, Z.: Gator: A unified framework for gaze object prediction. In: *CVPR*. pp. 19588–19597 (2022)

27. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Trans. Med. Imag.* **41**(7), 1688–1698 (2022)
28. Wenting, C., Jie, L., Yixuan, Y.: Bi-vlgm: Bi-level class-severity-aware vision-language graph matching for text guided medical image segmentation. *arXiv preprint arXiv:2305.12231* (2023)
29. Yang, X., Xie, J., Li, X., Li, X., Li, X., Shen, L., Deng, Y.: Tceip: Text condition embedded regression network for dental implant position prediction. In: *MICCAI*. pp. 317–326. Springer (2023)
30. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(12), 2878–2890 (2012)
31. Zhang, X., Wang, W., Chen, Z., Xu, Y., Zhang, J., Tao, D.: Clamp: Prompt-based contrastive learning for connecting language and animal pose. In: *CVPR*. pp. 23272–23281 (2023)