Gaze-DETR: Using Expert Gaze to Reduce False Positives in Vulvovaginal Candidiasis Screening

Yan Kong[1†], Sheng Wang[1,2†], Jiangdong Cai[1], Zihao Zhao[1], Zhenrong Shen[2], Yonghao Li[1], Manman Fei[2], and Qian Wang[1,3(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Clinical Research and Trial Center, Shanghai, China
qianwang@shanghaitech.edu.cn

Abstract. Accurate detection of vulvovaginal candidiasis is critical for women's health, yet its sparse distribution and visually ambiguous characteristics pose significant challenges for accurate identification by pathologists and neural networks alike. Our eye-tracking data reveals that areas garnering sustained attention - yet not marked by experts after deliberation - are often aligned with false positives of neural networks. Leveraging this finding, we introduce Gaze-DETR, a pioneering method that integrates gaze data to enhance neural network precision by diminishing false positives. Gaze-DETR incorporates a universal gaze-guided warm-up protocol applicable across various detection methods and a gaze-guided rectification strategy specifically designed for DETR-based models. Our comprehensive tests confirm that Gaze-DETR surpasses existing leading methods, showcasing remarkable improvements in detection accuracy and generalizability. Our code is available at https://github.com/YanKong0408/Gaze-DETR.

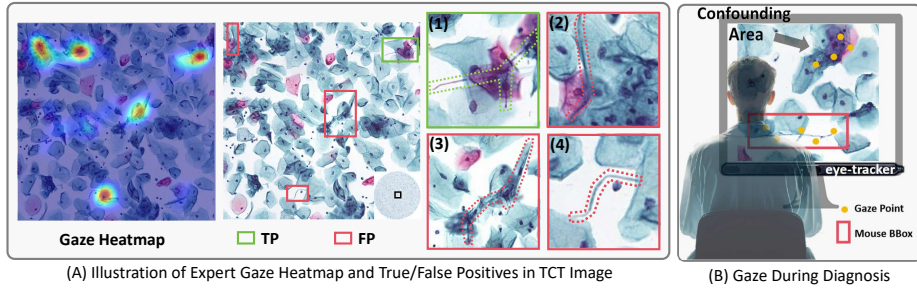**Keywords:** Candida detection, DETR, Eye-tracking

1 Introduction

Vulvovaginal candidiasis is a prevalent fungal infection caused by candida. It has been estimated that about 75% of women will experience vulvovaginal candidiasis at least once in their lifetime, leading to significant discomfort and healthcare resource usage [7,1,24]. Thin-prep cytologic test (TCT) plays a vital role to detect vulvovaginal candidiasis. However, visually analyzing TCT images requires intensive labor from pathologists, thus limiting the efficiency and scalability of cervical screening [10,3]. Therefore, there is an urgent need for a computer-assisted diagnosis (CAD) system capable of analyzing TCT images automatically.

In recent years, the emergence of deep-learning-based CAD systems has revolutionized cervical screening by facilitating automated detection of cervical abnormalities [17]. Despite notable progress, the existing methods still struggle with

---

† Yan Kong and Sheng Wang contributed equally to this study.

(A) Illustration of Expert Gaze Heatmap and True/False Positives in TCT Image   (B) Gaze During Diagnosis

Fig. 1. (A) TCT image samples displaying false positives (red) versus true positives (green) made by RetinaNet [12], and corresponding gaze heatmap. (B) Gaze tracking during labeling: An expert's gaze (yellow) and mouse-labeled regions (red). Regions where experts carefully review without annotating are usually confounding areas.

the precise identification of candida. The reasons include **a) the imbalance between candida and background cells**, and **b) the thin, faint appearance and frequent occlusions** as shown in Fig. 1(A) [3]. These factors hinder the model from distinguishing candida from similar cellular structures and contaminants, leading to a notable increase in false-positive results. An example TCT image is illustrated in Fig. 1(A), in which one true positive (green) and three false positives (red) with sharp cell structures and linear contaminants are displayed. It can be observed that RetinaNet [12] mistakenly identifies three (out of four) suspected regions as candida.

To address the issue of high false positive rates in TCT image analysis, we leverage the eye movement patterns of pathologists collected by eye-trackers during their image reading. The gaze data, which is long valued for enhancing diagnostic accuracy and understanding human error [2,19], has recently been shown to strengthen various CAD tasks [21,15,8,22,26,23,20]. By studying the connection between gaze patterns and detection errors, we observe a correlation between prolonged gaze on candida-like objects and the likelihood of those areas being identified as false positives by detection models. This insight, illustrated by a comparison between the detections of RetinaNet [12] and the expert gaze heatmap in Fig. 1(A), enables the strategic use of expert annotations and eye-tracking data to identify and correct potential false positives during model training. Note that the gaze data is passively collected when pathologists conduct manual labeling of candida boxes, causing no disruption to pathologists as depicted in Fig. 1(B).

With all the investigation, we propose **Gaze**-guided **De**tection **Tr**ansformer (Gaze-DETR), an innovative training strategy that incorporates expert gaze to enhance the ability of detectors to distinguish candida from false positives. Our model is built upon DETR [4] with a two-step training strategy comprising *gaze-guided warm-up* followed by *gaze-guided rectification*. Gaze-guided warm-up is employed to additionally predict the 'gaze only' category, highlighting regions where experts devote visual attention to carefully review yet without annotating
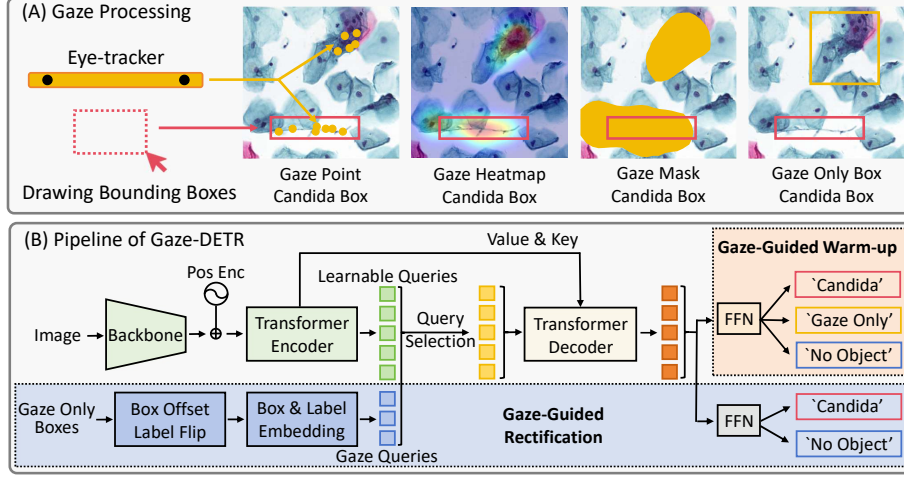
Fig. 2. Overview of the Gaze-DETR. (A) Illustration of Gaze Processing, which derives gaze points to 'gaze only' boxes. (B) Pipeline of the Gaze-DETR model, which integrates the gaze guidance into the DETR framework for candida detection.

them as actual candida instances. This warm-up strategy allows the detector to gain additional insights into confounding candida-like instances, thereby addressing the issue of imbalanced candida quantities. Subsequently, gaze-guided rectification is employed to construct more focused queries in the vicinity of confounding areas. These queries are then rectified by our model, which can be considered as a form of hard sample mining, effectively resolving the challenge of distinguishing difficult candida instances. The gaze information is not required during inference. The extensive experiments demonstrate the effectiveness of our approach in improving candida detection performance and its compatibility with various kinds of detection models.

2 Method

In this section, we start with the processing of gaze data, introducing the definition of 'gaze only' boxes along with their underlying semantics. Then, we delve into the pipeline of our proposed Gaze-DETR, which incorporates expert knowledge within 'gaze only' boxes via gaze-guided warm-up and gaze-guided rectification. By leveraging gaze data, we aim to align the recognition capabilities of DETR-like detectors with those of human experts, enhancing the overall performance of our Gaze-DETR model.

2.1 Gaze Processing

To mine semantics within gaze data that are beneficial to mitigate false negatives, we introduce the definition of 'gaze only' boxes and present its process pipeline in

Fig. 2(A). In this study, **'gaze only' boxes refer to regions receiving long-lasting attention from pathologists but are finally diagnosed as normal.** These 'gaze only' boxes are typically associated with instances (e.g. cellular folds and thin contaminants in Fig. 1(A)) that are challenging to discriminate in clinical practice, making them valuable for aligning DETR-like models with human recognition capabilities. To craft 'gaze only' boxes, a Gaussian kernel is first applied to those spatiotemporally discretized gaze points, generating a heatmap [21]. A threshold is then applied to this heatmap to identify regions of interest. Subsequently, the small-sized regions are eliminated, and gaze boxes are extracted by detecting contours within the remaining regions. Finally, the final 'gaze only' boxes are obtained by selecting the portions of the gaze boxes that do not contain candida boxes.

2.2   Gaze-Guided Warm-up

Compared with densely distributed cervical cells, the scarcity and sporadic distribution of candida in TCT images create an imbalance, which potentially leads to an unstable training phase. To overcome this, we propose a gaze-guided warm-up strategy for DETR-like detection models.

As shown in Fig. 2(B), DETR extracts features through a backbone and further forwards them through the Transformer encoder to obtain refined features. Unless otherwise specified, we follow the architecture of DAB-DETR [14] instead of the original DETR [4] throughout this study. Hence, learnable queries are learned from refind features, instead of random initialized and self-updated. These learnable queries are then processed by the Transformer decoder, finally obtaining predictions via the shared feed-forward networks (FFNs). To stabilize its training, we expand the FFN's classification capability beyond the 'candida' and 'no object' categories to also include a 'gaze only' category during the initial training phase. This strategy is plug-and-play across both DETR, and other CNN-based models as we will demonstrate in Section 3.3. While gaze only boxes are not actual candida, our strategy can 1) help the shallower layers of neural network learn to look at the discriminative appearance that looks like a candida. 2) prevent the deeper layers' gradient dilution when bias towards the majority (no object), which creates training instability.

2.3   Gaze-Guided Rectification

The matching strategy for 'no object' queries in DETR training significantly impacts model performance, especially in our task where a typical TCT image patch may contain only one or two instances of candida [25]. To enhance the model's focus on difficult areas, we introduce the strategy of gaze-guided rectification. This approach leverages gaze data to guide 'no object' queries towards confounding areas, providing targeted supervision during training and improving the model's discriminative ability. Fig. 3 contrasts the random distribution of 'no object' queries in standard DETR with our Gaze-DETR, which concentrates queries on expert-identified regions.
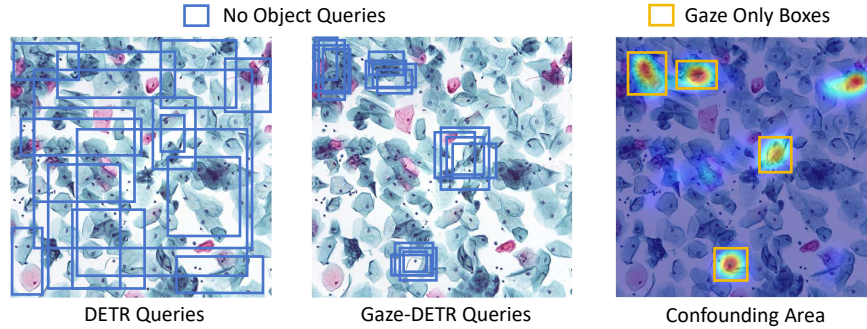
Fig. 3. Illustration to compare 'no object' queries of standard DETR against Gaze-DETR. Gaze-DETR effectively leverages 'gaze only' boxes to concentrate 'no object' queries on confounding regions, compared to the nearly random distribution of 'no object' queries of DETR.

To guide the detector's attention toward confounding areas, we introduce gaze queries, the position and content of which are both optimized by gaze data. Due to eye-tracking device errors and the gaze processing method, the generated 'gaze only' boxes tend to be offset from the corresponding objects, where the offset (Δx, Δy, Δw, Δh) follows respective Gaussian distributions. Thus, to make 'gaze only' boxes better reflect the confounding areas, we replicate the 'gaze only' boxes multiple times (four times in our model), and then utilize an explicit query position construction method, inspired by DAB-DETR [14], to offset the gaze boxes from (x', y', w', h') to (x'+ϵ_x, y'+ϵ_y, w'+ϵ_w, h'+ϵ_x), where ϵ_x, ϵ_y, ϵ_w, ϵ_h are random number following the corresponding Gaussian distribution. Additionally, we explicitly flip the class of the gaze boxes to candida, drawing inspiration from DN-DETR [11]. The class flip ensures that the class labels of the queries are corrected during processing in the FFNs, further prompting the model to match false positives to the 'no object' category. To maintain consistent token dimensions for the input of the decoder, we introduce a query selection mechanism. The mechanism prioritizes all gaze queries and selects the learnable queries with the most important features learned from the decoder as the rest. This mechanism enables the queries to concentrate on the confounding areas and candidate candida boxes to be retained.

3   Experimental Results

3.1   Data and Experimental Settings

Image Collection Our in-house dataset consists of whole slide images from patients with candida, which are cropped into patches of 1024×1024. These patches are meticulously examined and labeled by an experienced pathologist, who annotates candida with bounding boxes, resulting in a collection of 450
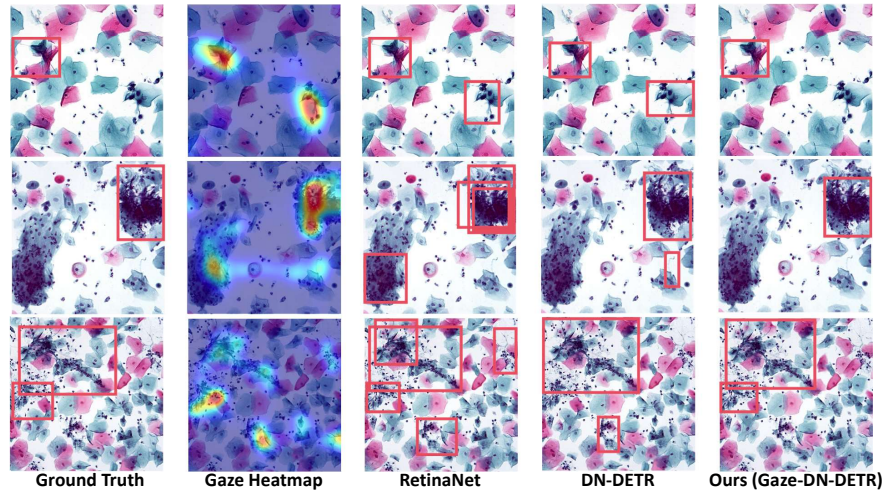
Fig. 4. Comparison of qualitative detection results. Our method significantly reduces false positive errors in areas that are carefully reviewed by experts, which are often prone to misclassifications by both RetinaNet and DN-DETR.

patches with 612 bounding boxes. The dataset is further divided into training, validation, and test sets with a ratio of 3:1:1.

Gaze Collection Gaze tracks were seamlessly obtained during the bounding box annotation process, utilizing software referenced in [21]. To enhance focus on cytology images, we refine the software by adding drag and zoom functionalities. The data collection is conducted across five distinct sessions, to ensure the accuracy of annotations and mitigate the impact of experimental fatigue.

Implementation Details To ensure fairness of the comparison, all experiments are conducted on one NVIDIA RTX 3090 GPU, with an initial learning rate of 3e-4. Our framework is optimized using Adam optimizer with cosine learning rate decay. CNN-based models are initialized with the default setting provided in MMDetection [6] and trained for 100 epochs. DETR-like models are initialized by the pre-train weights on CoCo dataset [13] for better performance [28] and trained for 50 epochs. Gaze-guided warmup is conducted in the early 10 epochs for CNN-based models and 5 epochs for DETR-like models. A quantitative evaluation is conducted using average precision (AP) and average recall (AR) as the metrics. We calculate the average AP over multiple IoU thresholds from 0.2 to 0.5 with a step size of 0.05 ($AP_{0.2:0.5}$), and individually evaluated AP at the IoU thresholds of 0.2 and 0.5 (denoted as $AP_{0.2}$ and $AP_{0.5}$), respectively.

3.2   Comparison with the State-of-the-Art

We first compare our method with the state-of-the-art detection algorithms to evaluate the improved performance using extra help from gaze. As shown in

Table 1. Comparison to state-of-the-art methods. Best results are denoted in bold.

Method	Backbone	$AP_{0.2:0.5}$	$AP_{0.2}$	$AP_{0.5}$	AR
EllipseNet (MICCAI 2021)[5]	ResNet-50	0.443	0.579	0.264	0.792
RetinaNet (ICCV 2017)[12]	ResNet-50	0.466	0.533	0.326	0.850
Sparse R-CNN (CVPR 2021) [18]	ResNet-50	0.386	0.553	0.149	0.685
YOLOv8 (2023)[9]	ResNet-50	0.482	0.587	0.333	0.866
DETR (ECCV 2020)[4]	ResNet-50	0.514	0.656	0.300	0.866
Conditional-DETR (ICCV 2021)[16]	ResNet-50	0.483	0.598	0.331	0.897
DN-DETR (CVPR 2022 Oral)[11]	ResNet-50	0.535	0.657	0.330	0.912
DINO (ICLR 2023)[25]	Swin-L	0.646	0.711	0.561	**0.988**
Gaze-DN-DETR	ResNet-50	0.573	0.697	0.348	0.893
Gaze-DINO	Swin-L	**0.687**	**0.755**	**0.634**	**0.988**

Table 1, DETR-like models generally show higher precision compared to CNN-based methods. Both DN-DETR and DINO exhibit improved performance with the assistance of gaze. Notably, Gaze-DINO achieves the best performance, which demonstrates the efficacy of Gaze-DETR in reducing false positives.

Qualitative comparisons are also conducted on methods with the backbone of ResNet-50. As shown in Fig. 4, both RetinaNet and DN-DETR tend to produce false positive results in areas where experts gaze but do not annotate bounding boxes. In contrast, our proposed method can significantly mitigate this situation with the help of expert gaze.

3.3 Compatibility of our method

To demonstrate the compatibility of our method and further validate its effectiveness, we extend our method to other detection models.

Gaze-guided warm-up can be integrated into CNN-based models seamlessly. A comparative analysis of RetinaNet and Sparse R-CNN, with and without Gaze-guided warm-up, is presented in the first two rows of Table 2. The results demonstrate that the incorporation of our method, specifically the inclusion of only gaze-guided warm-up, yields notable performance improvements in CNN-based models.

Any DETR-like model can utilize our proposed gaze integration framework. We conduct a comparison experiment on different DETR-like models with varying backbones. The latter part of Table 2 presents the results, from which several observations can be drawn: (1) The performance of DETR-like models consistently increases in most cases with the help of expert gaze. (2) Our proposed training strategy is also compatible with other techniques, such as denoise training [11], deformable attention [27], and query selection [25], etc.

3.4 Ablation Study

We conduct an ablation study on Gaze-DN-DETR with the ResNet-50 backbone and Gaze-DINO with the Swin-L backbone to verify the effectiveness of

Table 2. Results of extending our method to other detection models. Methods with * include only gaze-guided warm-up in the '+ Gaze' bar, while others include both gaze-guided warm-up and gaze-guided rectification.

Method	Backbone	Original				+ Gaze			
		$AP_{0.2:0.5}$	$AP_{0.2}$	$AP_{0.5}$	AR	$AP_{0.2:0.5}$	$AP_{0.2}$	$AP_{0.5}$	AR
RetinaNet*	ResNet-50	0.466	0.533	0.326	0.850	**0.478**	**0.542**	**0.332**	**0.873**
Sparse R-CNN*	ResNet-50	0.386	**0.553**	0.149	0.685	**0.394**	0.550	**0.168**	**0.692**
DN-DETR	ResNet-50	0.535	0.657	0.330	**0.912**	0.573	**0.697**	**0.348**	0.893
DN-deformable	ResNet-50	0.488	0.595	0.354	0.969	**0.502**	**0.609**	**0.367**	**0.979**
-DETR	ResNet-101	0.524	0.686	0.421	0.953	**0.568**	**0.721**	**0.440**	**0.976**
DINO	Resnet-50	0.481	0.576	**0.314**	0.941	**0.492**	**0.589**	0.310	**0.969**
	Swin-L	0.646	0.711	0.561	**0.988**	**0.687**	**0.755**	**0.585**	**0.988**

Table 3. Performance of ablation study for our Gaze-DN-DETR and Gaze-DINO. GGW denotes Gaze-Guided Warm-up and GGR denotes Gaze-Guided Rectification.

Method + GGW	+ GGR	DN-DETR (ResNet-50)				DINO (Swin-L)			
		$AP_{0.2:0.5}$	$AP_{0.2}$	$AP_{0.5}$	AR	$AP_{0.2:0.5}$	$AP_{0.2}$	$AP_{0.5}$	AR
		0.535	0.657	0.330	**0.912**	0.646	0.711	0.561	**0.988**
✓		0.565	0.685	0.330	0.880	0.672	0.749	0.584	0.981
	✓	0.532	0.669	0.339	0.902	0.661	0.751	0.554	0.960
✓	✓	**0.573**	**0.697**	**0.348**	0.893	**0.687**	**0.755**	**0.634**	**0.988**

each component and report the results in Table 3, from which several observations can be drawn: (1) Compared with the baseline, our gaze-guided warm-up achieves considerably better performance. (2) The gaze-guided rectification strategy alone leads to limited improvements. (3) Combining both gaze-guided warm-up and gaze-guided rectification strategies, our method demonstrates the best performance. This is because gaze-guided warm-up introduces an additional category, 'gaze only', in the early stages. This strengthens the classifier's capabilities and reduces the difficulty of gaze-guided rectification.

4   Conclusion

In this study, we propose Gaze-DETR, a novel training strategy that incorporates expert gaze patterns to reduce false positives in the identification of vulvovaginal candidiasis within TCT images. We seamlessly captured the eye movements of experts using an eye-tracker during data annotation. The gaze data offers a window into the diagnostic reasoning of pathologists by highlighting areas scrutinized but not ultimately annotated. We identified these areas as confounding regions that are prone to false positives for detectors. By implementing gaze-guided warm-up and gaze-guided rectification strategies, we effectively

harness this supplementary information to enhance the training of detection networks. Our experimental results affirm the efficacy and generalizability of Gaze-DETR, paving the way for more accurate medical image analysis through the integration of human expert insight.

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

References

1. Benedict, K., Jackson, B.R., Chiller, T., Beer, K.D.: Estimation of direct healthcare costs of fungal diseases in the united states. Clinical Infectious Diseases **68**(11), 1791–1797 (2019)
2. Brunyé, T.T., Drew, T., Weaver, D.L., Elmore, J.G.: A review of eye tracking for understanding and improving diagnostic interpretation. Cognitive research: principles and implications **4**(1), 1–16 (2019)
3. Cai, J., Xiong, H., Cao, M., Liu, L., Zhang, L., Wang, Q.: Progressive attention guidance for whole slide vulvovaginal candidiasis screening. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 233–242. Springer (2023)
4. 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. Springer (2020)
5. Chen, J., Zhang, Y., Wang, J., Zhou, X., He, Y., Zhang, T.: Ellipsenet: anchor-free ellipse detection for automatic cardiac biometrics in fetal echocardiography. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. pp. 218–227. Springer (2021)
6. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
7. Gonçalves, B., Ferreira, C., Alves, C.T., Henriques, M., Azeredo, J., Silva, S.: Vulvovaginal candidiasis: Epidemiology, microbiology and risk factors. Critical reviews in microbiology **42**(6), 905–927 (2016)
8. Ji, C., Du, C., Zhang, Q., Wang, S., Ma, C., Xie, J., Zhou, Y., He, H., Shen, D.: Mammo-net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 68–78. Springer (2023)
9. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Jan 2023)
10. Koss, L.G.: The papanicolaou test for cervical cancer detection: a triumph and a tragedy. Jama **261**(5), 737–743 (1989)
11. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
14. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
15. Ma, C., Zhao, L., Chen, Y., Wang, S., Guo, L., Zhang, T., Shen, D., Jiang, X., Liu, T.: Eye-gaze-guided vision transformer for rectifying shortcut learning. IEEE Transactions on Medical Imaging (2023)
16. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3651–3660 (2021)
17. Rahaman, M.M., Li, C., Wu, X., Yao, Y., Hu, Z., Jiang, T., Li, X., Qi, S.: A survey for cervical cytopathology image analysis using deep learning. IEEE Access **8**, 61687–61710 (2020)
18. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14454–14463 (2021)
19. Voisin, S., Pinto, F., Morin-Ducote, G., Hudson, K.B., Tourassi, G.D.: Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography. Medical physics **40**(10), 101906 (2013)
20. Wang, B., Pan, H., Aboah, A., Zhang, Z., Keles, E., Torigian, D., Turkbey, B., Krupinski, E., Udupa, J., Bagci, U.: Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2194–2203 (2024)
21. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)
22. Wang, S., Zhao, Z., Zhang, L., Shen, D., Wang, Q.: Crafting good views of medical images for contrastive learning via expert-level visual attention. In: NeuRIPS 2023 Workshop on Gaze Meets ML (2023)
23. Wang, S., Zhuang, Z., Ouyang, X., Zhang, L., Li, Z., Ma, C., Liu, T., Shen, D., Wang, Q.: Learning better contrastive view from radiologist's gaze. arXiv preprint arXiv:2305.08826 (2023)
24. Willems, H.M., Ahmed, S.S., Liu, J., Xu, Z., Peters, B.M.: Vulvovaginal candidiasis: a current understanding and burning questions. Journal of Fungi **6**(1), 27 (2020)
25. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
26. Zhao, Z., Wang, S., Wang, Q., Shen, D.: Mining gaze for contrastive learning toward computer-assisted diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7543–7551 (2024)
27. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
28. Zhu, Y., Shen, Z., Zhao, Z., Wang, S., Wang, X., Zhao, X., Shen, D., Wang, Q.: Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis. arXiv preprint arXiv:2311.08236 (2023)