Predicting Radiologist Expertise from 3D Gaze Patterns During CT Interpretation

Leila Khaertdinova[1,*], Anna Anikina[1,*], Claudia Mello-Thoms[2], and Bulat Ibragimov[1]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
{leila.khaertdinova, anan, bulat}@di.ku.dk
[2] Department of Radiology, University of Iowa, Iowa, United States
claudia-mello-thoms@uiowa.edu

Abstract. Accurate interpretation of volumetric CT requires efficient navigation of 3D image volumes and attention to diagnostically relevant regions. While eye-tracking has been widely studied in 2D medical imaging, its use for expertise assessment in CT settings remains limited. We propose a gaze-informed transformer framework for expertise classification in thoracic CT. Using a DINOv2 backbone, radiologist fixation patterns are integrated into volumetric feature learning through (1) a learnable log-space bias in self-attention and (2) gaze-weighted pooling of patch embeddings. We trained and evaluated our approach on 182 CT reading sessions from five radiologists with varying levels of experience. On a held-out test set, the model achieves an ROC-AUC of 0.91 and F1 score of 0.86, outperforming adapted methods. These findings suggest that incorporating visual search behavior into transformers may support objective, process-based expertise assessment in radiology. Code is available via https://github.com/leiluk1/GazeToSkill.

Keywords: Eye tracking · Skill assessment · Vision Transformer

1 Introduction

Accurate and objective assessment of clinical expertise is essential for developing training programs that promote the adoption of expert-like strategies and accelerate skill acquisition [3]. In radiology, expertise is reflected in visual search behavior, with consistent differences reported between experts and novices [4]. Capturing these differences requires instrumentation capable of tracking spatial and temporal patterns of attention during diagnostic reading, and eye-tracking is one of the tools that enables quantitative measurement of visual behavior [3,4]. However, reported gaze-based differences are often task- and modality-dependent, indicating that handcrafted statistical measures may not fully capture the complexity of visual expertise [7,8].

[*] These authors contributed equally to this work.

Deep learning (DL) offers a complementary approach for modeling high-dimensional spatiotemporal gaze patterns, potentially revealing subtle structures associated with expertise [12]. Most existing studies dedicated to gaze-based skill assessment typically operate on 2D views or video clips and do not explicitly model the three-dimensional spatial context of volumetric imaging [1,14,18,22]. Therefore, it remains unclear how well these approaches generalize to 3D CT settings. A related line of research within this domain investigates gaze scanpath prediction, where DL models are trained on expert gaze recordings to learn sequential patterns of visual search. After training, these models can generate expert-like gaze trajectories conditioned on input medical images, providing an implicit representation of expert attention strategies, and potentially distinguish experts' scanpaths from novices. However, to the best of our knowledge, only one study has investigated scanpath prediction for 3D CTs [19].

Motivated by the limited exploration of expertise modeling in 3D medical imaging, we aim to enrich the field of gaze-based skill assessment by investigating expert attention patterns in volumetric CT. In clinical training, such a system can support supervision by providing an estimate of reader expertise from a single CT interpretation session. Novice-like predictions across readings may flag incomplete acquisition of expert search strategies and help to identify individuals who require additional training [3]. We introduce a gaze-informed transformer framework that integrates radiologist visual attention directly into volumetric CT representation learning. Using a DINOv2 vision transformer (ViT), we inject per-patch gaze signals at two complementary levels: (1) within the self-attention mechanism via a learnable log-space bias that multiplicatively reweights attention probabilities, and (2) at the representation level through gaze-weighted pooling of patch embeddings. The additive logit bias increases the influence of highly fixated regions during attention computation while still allowing the transformer to model global context. This design enables the model to learn not only *what* is visible in a CT scan, but also *where* expert radiologists allocate their attention. Furthermore, we collected 182 CT reading sessions with synchronized eye-tracking data from five expert and novice radiologists and compared our methodology against adapted approaches on our dataset.

2 Dataset

2.1 Data Collection

CT Dataset. Three expert radiologists (7+ years of professional experience) and two novice readers (<1 year of experience) analyzed 40 lung CT scans sourced from the publicly available LIDC–IDRI dataset [2]. Among the images, 24 scans corresponded to lung cancer patients, while the remaining 16 were from non-cancer patients. The cancer cases contained 47 individual lung nodules, including 20 small-to-intermediate–sized nodules [21] and 27 large nodules [9]. Each reader's diagnostic performance for cancer detection was assessed based on the synchronized gaze and audio recordings: experts reached a mean sensitivity

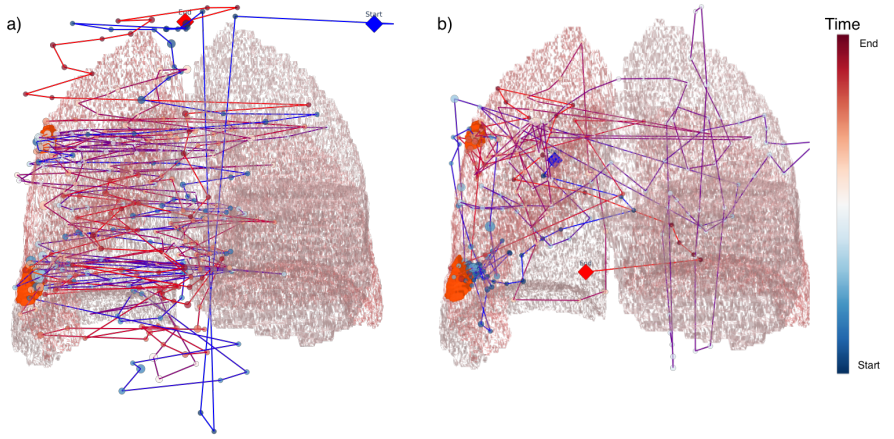

Fig. 1. Example scanpaths from an expert (a) and a novice (b). The trajectories indicate the sequence of fixations over time, with the time bar normalized. Lung nodules are indicated by orange segmentation masks.

of 0.96 and specificity of 0.98, while novices reached 0.89 and 0.58, respectively, supporting the assigned expertise levels.

Experimental Protocol. Gaze data were recorded with a Tobii Eye Tracker 4C (90 Hz) mounted below a 23.7-inch 4K 10-bit monitor, while audio reports were captured via a headset microphone. Radiologists completed calibration before the experiment, with recalibration performed after posture changes. CT scans were reviewed in the RadiAnt DICOM Viewer [16]. Recordings were started and ended verbally ("START"/"END") and synchronized by fixating on a predefined red point while stating "RED POINT." Radiologists followed their routine clinical workflow, including zooming, adjusting window level (WL) and window width (WW), and navigating across different orthogonal planes.

2.2 Data Preprocessing

Gaze-to-frame Synchronization. Raw eye-tracking coordinates were synchronized with video frames, so only frames with gaze data were processed. Gaze drift was corrected by computing the deviation from a reference red point at the start and end of each session and applying time-based linear interpolation between these two data points. In addition, gaze drift was adjusted using the nodule location. For each frame, we used EasyOCR [13] to automatically extract metadata displayed in the viewer, including CT plane, slice number, WL, WW.

Gaze-to-CT Coordinate Mapping. CT slice numbers were used to retrieve the corresponding slices from the original NIfTI volume and windowed using the extracted WL/WW parameters. Each slice was aligned to the video frame via affine transformation (accounting for scaling, rotation, and translation), and gaze coordinates were mapped from screen space to CT image space using the inverse affine matrix, yielding pixel-accurate locations in the original CT volume.

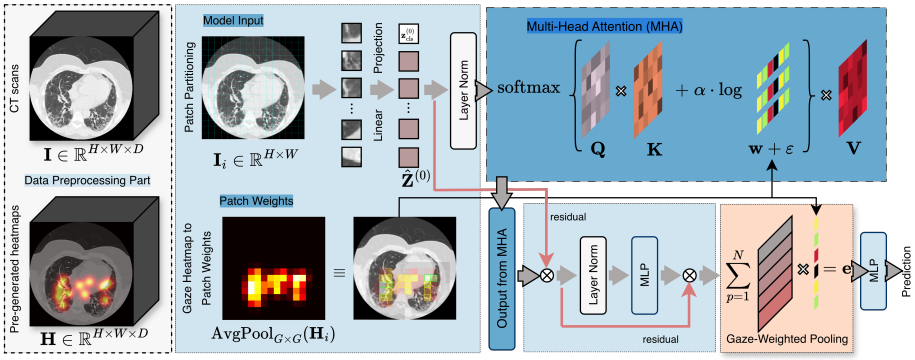

Fig. 2. Overview of the proposed gaze-informed transformer model. The input CT volume is partitioned into patches and projected into tokens, while the heatmap volume is downsampled and normalized to obtain per-patch gaze weights. The weights are injected into each self-attention layer as a log-space bias, reweighting attention toward fixated regions. Final slice embeddings are obtained via gaze-weighted pooling, aggregated across slices, and fed to a MLP classifier that predicts the expertise class.

Fixation Extraction. Fixations were classified using the Dispersion-Threshold Identification (I-DT) method [20], which groups spatially clustered gaze points that persist over time. The duration threshold was set to 100 ms and the dispersion threshold to 1° of visual angle, as recommended in the literature [11,20].
Heatmap Generation. We generated fixation heatmaps for each slice by smoothing projected gaze points with a Gaussian kernel corresponding to 1° of visual angle. The Gaussian standard deviation was computed per session using the mean viewing distance and the screen-to-CT pixel ratio, enabling subject-specific spatial calibration. The resulting heatmaps were normalized.

3 Methods

Input. The full CT volume is defined as $\mathbf{I} \in \mathbb{R}^{H\times W\times D}$, where D is a number of axial slices. For each slice $\mathbf{I}_i \in \mathbb{R}^{H\times W}$, WW and WL are applied. The greyscale slice is triplicated to form a pseudo-RGB image $\mathbf{X}_i \in \mathbb{R}^{3\times H'\times W'}$.
Gaze Heatmap to Patch Weights. Each radiologist's gaze session has a heatmap volume $\mathbf{H} \in \mathbb{R}^{H\times W\times D}$. For slice i, heatmap $\mathbf{H}_i \in \mathbb{R}^{H\times W}$ is downsampled to the patch grid: $\tilde{\mathbf{W}}_i = \mathrm{AvgPool}_{G\times G}(\mathbf{H}_i) \in \mathbb{R}^{G\times G}$, where $G = H'/P$ is the patch grid size and P is the ViT patch size. The pooled map is flattened and ℓ_1-normalized to obtain per-patch gaze weights $\mathbf{w}$:

$$\tilde{w}_{i,p} = \mathrm{vec}(\tilde{\mathbf{W}}_i)_p, \quad p = \overline{1,N}, \quad N = G^2. \tag{1}$$

$$w_{i,p} = \begin{cases} \dfrac{\tilde{w}_{i,p}}{\sum_{j=1}^{N} \tilde{w}_{i,j}} & \text{if } \sum_{j=1}^{N} \tilde{w}_{i,j} > 0, \\ \dfrac{1}{N} & \text{otherwise.} \end{cases} \tag{2}$$

Patch Embedding. Each slice $\mathbf{X}_i \in \mathbb{R}^{3\times H'\times W'}$ is partitioned into non-overlapping patches of size $P \times P$ and linearly projected to obtain N patch tokens $\mathbf{z}_p^{(0)}, p = \overline{1, N}$. A learnable classification token (CLS) $\mathbf{z}_{\text{cls}}^{(0)} \in \mathbb{R}^{D_e}$, where D_e is embedding dimension, is prepended, and learnable positional embeddings are added. The resulting sequence $\hat{\mathbf{Z}}^{(0)}$ is passed through L transformer blocks.

Gaze-Bias Attention. At each transformer layer $\ell = \overline{1, L}$, input $\hat{\mathbf{Z}}^{(\ell-1)}$ is transformed into queries, keys, and values via learned projections $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$. Gaze is injected as an additive log-space bias derived from per-patch gaze weights:

$$A_{ij}^{\text{gaze}} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_h}} + \alpha \cdot \log(w_j + \varepsilon), \tag{3}$$

where $d_h = D_e/n_h$ is the per-head dimension and n_h is the number of attention heads, α is a learnable scalar, ε is a small constant that prevents numerical underflow, and w_j denotes the gaze weight for key position j. The CLS key receives zero bias by setting $w_0 = 1$. The bias is shared across all heads within a layer and is identical across all L layers. Applying the exponential to the gaze-biased logit yields:

$$\text{softmax}(z_j + \alpha \log w_j) \propto e^{z_j} \cdot e^{\alpha \log w_j} = e^{z_j} \cdot w_j^\alpha, \tag{4}$$

where $z_j = \frac{q_i \cdot k_j}{\sqrt{d_h}}$ denotes the standard attention logit. Thus, the additive log-space bias in logit space becomes a multiplicative scaling in probability space: each key's attention weight is multiplied by w_j^α, amplifying heavily fixated patches and suppressing those that received little gaze. Then, the output of each transformer block follows the standard residual structure.

Gaze-Weighted Output Pooling. Given the final patch tokens $\mathbf{z}_p \in \mathbb{R}^{D_e}$ and normalized gaze weights w_p for slice i, the slice-level embedding is computed as the weighted average:

$$\mathbf{e}_i = \sum_{p=1}^{N} w_{i,p}\, \mathbf{z}_p. \tag{5}$$

This directs the representation towards image regions that received visual attention, encoding where the radiologist looked during interpretation.

Slice Sampling. During training, a subset of K slices is uniformly sampled from the D available slices. After Gaze-Weighted Output Pooling, the slice-level embeddings are aggregated by mean pooling across all K slices: $\bar{\mathbf{e}} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{e}_i$.

Classification Head. The session embedding $\bar{\mathbf{e}} \in \mathbb{R}^{D_e}$ is passed through a two-layer MLP:

$$\hat{\mathbf{y}} = \mathbf{W}_2\, \text{Dropout}(\text{ReLU}(\mathbf{W}_1 \bar{\mathbf{e}} + \mathbf{b}_1)) + \mathbf{b}_2, \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D_h \times D_e}$, $\mathbf{W}_2 \in \mathbb{R}^{2\times D_h}$, and D_h is the hidden dimension. The output $\hat{\mathbf{y}} \in \mathbb{R}^2$ represents logits for the novice and expert classes.

4 Results

Dataset. Our dataset comprises 40 CTs with 8,022 axial slices (∼200 slices per volume). Across all radiologists, sessions lasted 117.9±60.3 s and contained 160± 78 fixations (mean duration 0.48 ± 0.25 s, cumulative dwell time 73.1 ± 38.0 s).
Implementation Details. After excluding incomplete recordings, 182 samples remained (3 experts: A, B, C and 2 novices: D, E). Radiologists C and E were held out for testing (67 samples: 39 expert, 28 novice), while the remaining 115 samples were used for training. All models were trained using stratified 5-fold cross-validation to address class imbalance (∼70% expert). At inference phase, fold predictions were averaged, and the final classification threshold was determined using Youden's J on the test set. For *Skill Assessment Models* as well as for our approach, the training set included 80 expert and 35 novice samples. *Adapted Scanpath Prediction Models* were trained only on expert data (A and B; 80 samples), and expertise classification was based on similarity metrics.

For training our model, we used Adam with a learning rate of $1e^{-5}$ for the classification head and $1e^{-6}$ for the backbone. Each training step sampled 8 random slices per CT volume with gradient accumulation over 4 steps, yielding a batch size of 4 sessions. Models were trained for 100 epochs with the best checkpoint selected by validation ROC-AUC. The classification head consisted of two linear layers ($768 \rightarrow 256 \rightarrow 2$) with ReLU and dropout (p=0.1), trained with binary cross-entropy (BCE) loss. Input CT slices were windowed to lung settings (WL=−600, WW=1500) and resized to 518×518. For the DINOv2 ViT-B/14 backbone, the patch size was 14×14, and the embedding dimension was 768. The model had 86.8 M trainable parameters and was implemented using PyTorch Lightning v2.6.0 with PyTorch 2.7.0 (CUDA 12.6) and trained on an NVIDIA L40S GPU for approximately 17 hours.
Evaluation. We compared our approach to two categories of existing models, adapted and trained on our CT data (see Table 1).
Skill Assessment Models: We adapted four multimodal CNN architectures with strong reported performance, originally proposed by Sharma et al. [22] for fetal ultrasound skill classification: Late Fusion (LF-CNN), Intermediate Fusion (IF-CNN), Hybrid Fusion (HF-CNN), and Tensor Fusion (TF-CNN). Original input modalities included Standard Plane (SP), Spatial Gaze Maps (SGP), Gaze Trajectory Images (GTI), and Pupillary Response Images. Pupillary Response Images were omitted due to unavailability; and SP images are operator-acquired anatomical views that reflect skill level, instead we use fixation-ordered CT scans augmented with gaze information (FO-CT+Gaze), where expertise is expressed through slice navigation behavior; generation of SGM and GTI repeated original pipeline [22]. We separately tested two modalities (SGM and GTI) and three modalities (FO-CT+Gaze, SGM, GTI) to show contribution of SP replacement.
Adapted Scanpath Prediction Models: We trained two models on expert gaze data to predict scanpaths; the similarity between a reader's actual scanpath and the model prediction can thus serve as an indicator of expertise. Lou et al. introduced a multi-stream model with three backbones (ConvNeXt-B, HRNet-W48, CSwin Transformer) that achieved state-of-the-art gaze saliency prediction for mammo-

Table 1. Comparison against prior methods on the held-out test set using ROC-AUC, F1, Sensitivity, and Specificity metrics. SGP: Spatial Gaze Map; GTI: Gaze Trajectory Image; FO-CT+Gaze: fixation-ordered CT slices augmented with gaze information.

Model	Input	Prediction	ROC-AUC	F1	Sens.	Spec.
TF-CNN [22]	SGP, GTI	Skill	0.7793	0.8372	**0.9231**	0.6071
	FO-CT+Gaze, SGP, GTI		0.7454	0.7568	0.7179	0.7500
IF-CNN [22]	SGP, GTI	Skill	0.7308	0.6866	0.5897	0.8214
	FO-CT+Gaze, SGP, GTI		0.7363	0.7532	0.7436	0.6786
LF-CNN [22]	SGP, GTI	Skill	0.7463	0.7838	0.7436	0.7857
	FO-CT+Gaze, SGP, GTI		0.7637	0.8205	0.8205	0.7500
HF-CNN [22]	SGP, GTI	Skill	0.7518	0.7838	0.7436	0.7857
	FO-CT+Gaze, SGP, GTI		0.7363	0.7733	0.7436	0.7500
Lou et al. [15]	FO-CT session-level	Saliency map	0.6782	0.7123	0.6667	0.7241
	FO-CT chunk-level		0.5570	0.6000	0.5385	0.6552
CT-Searcher [19]	CT volume	Scanpath	0.8750	0.8421	0.8205	0.8214
Ours	CT volume	Skill	**0.9089**	**0.8611**	0.7949	**0.9310**

grams [15]. We adapted their framework from 2D mammography to chest CT by projecting 3D gaze data into 2D using Maximum Intensity Projection (MIP). We evaluated two MIP strategies: (1) a session-level MIP, collapsing all viewed slices per reading into a single saliency map, and (2) a chunk-level MIP, splitting each reading into groups of 20 consecutive slices to preserve local depth context. We then thresholded the skill prediction based on Normalized Scanpath Saliency (NSS) metric. CT-Searcher [19], a transformer-based scanpath predictor, was trained on the CTScanGaze dataset [19] and fine-tuned on our expert data. The predictions were then thresholded based on MultiMatch vector similarity [6].

Ablation study. We evaluated DINOv2 [17], Med3D [5], SwinUNETR [10], and UniMISS [23] for CT slice embeddings. Med3D [5] showed representational collapse (cross-CT 0.9998; variance 0.0002), indicating minimal discriminative capacity; UniMISS [23] demonstrated weak inter-patient separation (0.9801 ± 0.0193). SwinUNETR [10] improved between-patient discrimination (0.9349 ± 0.0668) but exhibited near-saturated consecutive similarity (0.9999), suggesting limited sensitivity to subtle slice-level variations. DINOv2 [17] achieved the best trade-off, maintaining clear inter-patient separation (0.9381 ± 0.0261) with higher representational variance (0.6814), and was therefore selected as a backbone.

To assess individual contributions of gaze-integration mechanisms, we conducted an ablation study (Table 2). Firstly, we trained MLP with frozen DINOv2 embeddings and obtained near-random results for prediction (Mean Val AUC: 0.4455 ± 0.045, AUC on test: 0.5234). Then, we unfroze DINOv2 and tested configurations over two axes: the attention mode (Gaze-Bias, Fixation-Mask, or None) and pooling strategy (Gaze-Weighted or CLS).

Table 2. Attention variants: Gaze-Bias - soft additive bias injected into self-attention layers; Fix-Mask - hard binary mask on attention, meaning patches with zero gaze weight are blocked from being attended to; None - standard DINOv2 self-attention, no gaze signal enters the transformer. Pooling strategies: Gaze-Weighted - weighted average of all patch tokens using normalized gaze weights, i.e., patches the radiologist looked at more contribute more to the session embedding; CLS - uses only CLS token.

Attention	Pooling strategies	Mean Val AUC	AUC	F1	Sens.	Spec.
Gaze-Bias	Gaze-Weighted	0.796 ± 0.091	**0.909**	**0.861**	**0.795**	0.931
Gaze-Bias	CLS	0.828 ± 0.101	0.897	0.824	0.751	0.892
None	Gaze-Weighted	0.632 ± 0.090	0.856	0.759	0.623	**0.969**
Fix-Mask	Gaze-Weighted	0.871 ± 0.101	0.842	0.783	0.688	0.892

5 Discussion and Conclusion

According to Table 1, our proposed model achieved the best performance across almost all metrics, with the highest AUC of 0.91, F1 of 0.86, and specificity of 0.93. The multi-modal fusion architectures introduced by Sharma et al. [22], originally designed for ultrasound, achieved comparable performance when adapted to CT, indicating that gaze patterns reflect general expertise rather than task-specific behaviors. Among these, TF-CNN with SGP, GTI inputs achieved the highest sensitivity (0.92), though with lower specificity (0.61), indicating that novices were frequently misclassified as experts. The addition of CT image information (FO-CT+Gaze) did not consistently improve performance across architectures, with IF-CNN showing little improvement (ROC-AUC: 0.73 → 0.74) and LF-CNN (ROC-AUC: 0.75 → 0.76). This suggests that the FO-CT+Gaze representation may provide limited additional value beyond other features, or that the increased model complexity leads to overfitting given the dataset size. Lou et al. [15] relies on a three-encoder architecture that is computationally expensive in both time and memory, limiting its scalability to volumetric data. Chunk-level processing, necessary to reduce computational demands, achieved the ROC-AUC of only 0.56, suggesting that this architecture may be difficult to adapt fully for 3D CT interpretation tasks. CT-Searcher [19], originally designed for scanpath task, achieved accurate results (ROC-AUC 0.88) despite not being explicitly designed for skill classification, indicating that learned representations for scanpath modeling can encode skill-relevant information.

The ablation study, provided in Table 2, reveals that gaze information contributes to expertise classification through two complementary mechanisms. The best-performing configuration (Gaze-Bias + Gaze-weighted, AUC = 0.91) integrates gaze at both the attention and pooling levels, substantially outperforming variants where gaze operates at only one stage: either biasing attention alone (Gaze-Bias + CLS, AUC = 0.89) or weighting pooling alone (None + Gaze-weighted, AUC = 0.86). This suggests that gaze serves two distinct roles: during feature extraction, it guides the model to focus its internal processing on regions the radiologist actually examined, while during aggregation, it ensures that heavily fixated areas contribute more to the final classification. In other words, gaze

tells the model both how to analyze each slice and what matters most within it, and both signals are needed to distinguish experts from novices.

The clinical significance of our work lies in enabling objective image reading assessment of radiologist expertise. This has practical implications for radiology training, where automated gaze-based feedback could identify specific visual weaknesses in trainees. Understanding which gaze behaviors reliably mark expertise may also inform the design of computer-aided detection systems tailored to the perceptual gaps of less experienced readers. Future work will focus on increasing the number of participants from five radiologists to support more extensive clinical validation.

Acknowledgments. This work was supported by the Novo Nordisk Foundation under Grant NFF20OC0062056, and the European Research Council under Grant AI-Dose:101171810.

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

References

1. Akerman, M., Choudhary, S., Liebmann, J.M., Cioffi, G.A., Chen, R.W., Thakoor, K.A.: Extracting decision-making features from the unstructured eye movements of clinicians on glaucoma oct reports and developing ai models to classify expertise. Frontiers in Medicine **10**, 1251183 (2023)
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: Data from LIDC-IDRI [Data set]. The Cancer Imaging Archive (2015)
3. Ashraf, H., Sodergren, M.H., Merali, N., Mylonas, G., Singh, H., Darzi, A.: Eye-tracking technology in medical education: A systematic review. Medical teacher **40**(1), 62–69 (2018)
4. 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), 7 (2019)
5. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
6. Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., Holmqvist, K.: It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. Behavior research methods **44**(4), 1079–1100 (2012)
7. Gegenfurtner, A., Lehtinen, E., Säljö, R.: Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. Educational psychology review **23**(4), 523–552 (2011)
8. Van der Gijp, A., Ravesloot, C., Jarodzka, H., Van der Schaaf, M., Van der Schaaf, I., Ten Cate, T.J.: How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education **22**(3), 765–787 (2017)
9. Han, D., Heuvelmans, M.A., Oudkerk, M.: Volume versus diameter assessment of small pulmonary nodules in ct lung cancer screening. Translational lung cancer research **6**(1), 52 (2017)

10. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
11. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye tracking: A comprehensive guide to methods and measures. oup Oxford (2011)
12. Ibragimov, B., Mello-Thoms, C.: The use of machine learning in eye tracking studies in medical imaging: A review. IEEE journal of biomedical and health informatics **28**(6), 3597–3612 (2024)
13. JaidedAI: EasyOCR: Ready-to-use OCR with 80+ supported languages, `https://github.com/JaidedAI/EasyOCR`, last accessed 2025/01/30
14. Lam, K., Chen, J., Wang, Z., Iqbal, F.M., Darzi, A., Lo, B., Purkayastha, S., Kinross, J.M.: Machine learning for technical skill assessment in surgery: a systematic review. NPJ digital medicine **5**(1), 24 (2022)
15. Lou, J., Lin, H., Young, P., White, R., Yang, Z., Shelmerdine, S., Marshall, D., Spezi, E., Palombo, M., Liu, H.: Predicting radiologists' gaze with computational saliency models in mammogram reading. IEEE Transactions on Multimedia **26**, 256–269 (2023)
16. Medixant: Radiant DICOM Viewer, `https://www.radiantviewer.com`, last accessed 2025/06/01
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
18. Pedrett, R., Mascagni, P., Beldi, G., Padoy, N., Lavanchy, J.L.: Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review. Surgical endoscopy **37**(10), 7412–7424 (2023)
19. Pham, T.T., Awasthi, A., Khan, S., Marti, E.D., Nguyen, T.P., Vo, K., Tran, M., Nguyen, S., Tran, C., Ikebe, Y., et al.: Ct-scangaze: A dataset and baselines for 3d volumetric scanpath modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21732–21743 (2025)
20. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 symposium on Eye tracking research & applications. pp. 71–78 (2000)
21. Sánchez, M., et al.: Management of incidental lung nodules< 8 mm in diameter. Journal of thoracic disease **10**(Suppl 22), S2611 (2018)
22. Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Multi-modal learning from video, eye tracking, and pupillometry for operator skill characterization in clinical fetal ultrasound. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1646–1649. IEEE (2021)
23. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)