This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CT-ScanGaze: A Dataset and Baselines for 3D Volumetric Scanpath Modeling

Trong Thang Pham[1], Akash Awasthi[2], Saba Khan[2], Esteban Duran Marti[1],
Tien-Phat Nguyen[3], Khoa Vo[1], Minh Tran[1], Son Nguyen[4], Cuong Tran[4], Yuki Ikebe[1],
Anh Totti Nguyen[5], Anh Nguyen[6], Zhigang Deng[2], Carol C. Wu[7], Hien Nguyen[2], and Ngan Le[1]

[1]University of Arkansas, [2]University of Houston, [3]VNUHCM John von Neumann,
[4]FPT Software AI Center, [5]Auburn University, [6]University of Liverpool, [7]MD Anderson Cancer Center

Abstract

Understanding radiologists' eye movement during Computed Tomography (CT) reading is crucial for developing effective interpretable computer-aided diagnosis systems. However, CT research in this area has been limited by the lack of publicly available eye-tracking datasets and the three-dimensional complexity of CT volumes. To address these challenges, we present the first publicly available eye gaze dataset on CT, called CT-ScanGaze, captured from expert radiologists. Then, we introduce CT-Searcher, a novel 3D scanpath predictor designed specifically to process CT volumes and generate radiologist-like 3D fixation sequences, overcoming the limitations of current scanpath predictors that only handle 2D inputs. Since deep learning models benefit from a pretraining step, we develop a pipeline that converts existing 2D gaze datasets into 3D gaze data to pretrain CT-Searcher. Through both qualitative and quantitative evaluations on CT-ScanGaze, we demonstrate the effectiveness of our approach and provide a comprehensive assessment framework for 3D scanpath prediction in medical imaging. Code and data are available at https://github.com/UARK-AICV/CTScanGaze.

1. Introduction

Interpretability is a fundamental aspect of Computer-aided Diagnosis (CAD) system as it supports safe, accurate, and trustworthy patient care [47]. The integration of eye gaze signals offers a promising approach to enhance the interpretability of CAD systems [22, 38, 47, 48, 51]. Understanding the importance of gaze data in medical imaging analysis, EGD [38] and REFLACX [6] are created and shared publicly to advance chest X-rays (CXR) analysis. As shown in Fig. 1a, these datasets have facilitated developments for artificial intelligent solutions, including multi-label classification [38, 62, 69], report generation [38, 61], explainable classification [55], explainable report generation [54], and scanpath prediction [67]. However, existing medical datasets are either limited to 2D images like CXRs or lack radiologist gaze data. The absence of public dataset capturing radiologists' eye gaze patterns on Computed Tomography (CT) has left several critical tasks underexplored in CT imaging, including scanpath prediction, explainable

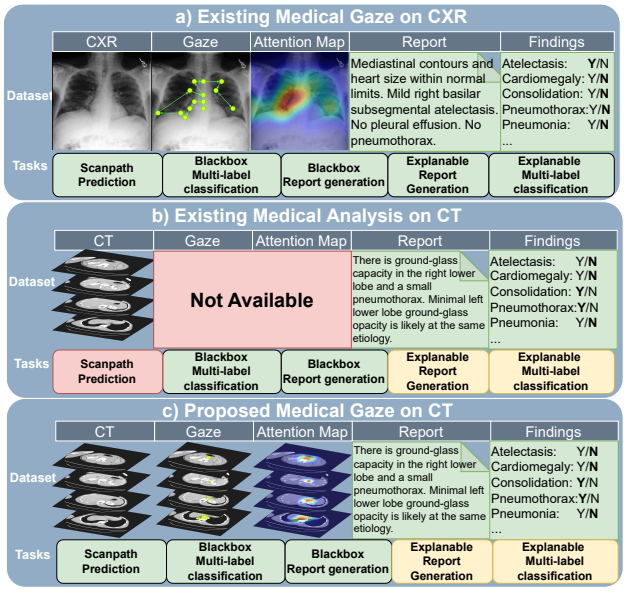


Figure 1. Many research directions in CAD would benefit from the availability of gaze, report, and findings data, as is the case for CXR (a), highlighted in green. However, some critical areas in CT research are underexplored. For example, there are only preliminary results for Explainable Report Generation and Explainable Classification, highlighted in yellow. Especially, Scanpath Prediction is underexplored and often overlooked, highlighted in red, primarily due to the lack of publicly available datasets (b). Our dataset offers new research opportunities to these tasks. In this paper, we address the Scanpath Prediction task on CT scans (c).

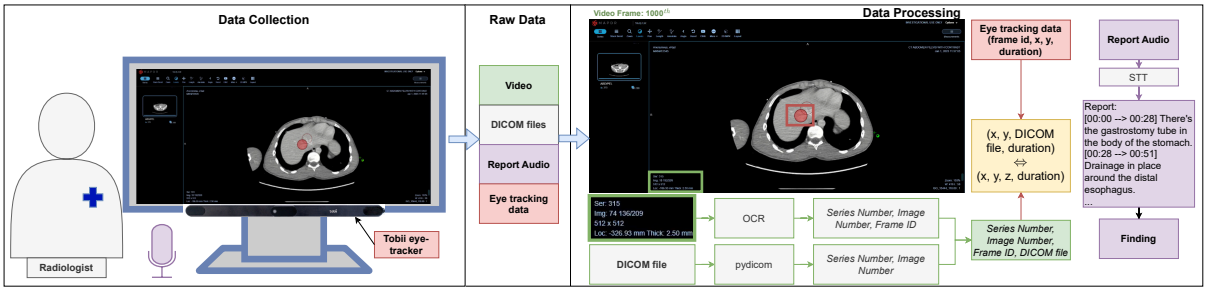


Figure 2. Illustration of our data collection and processing pipeline. The Data Collection panel shows the setup: a radiologist examines CT scans on a monitor equipped with a Tobii eye-tracker and microphone for recording audio reports. The Raw Data panel displays four data streams collected: screen recording video, DICOM files, verbal report audio, and eye gaze data. The Data Processing panel demonstrates the data processing pipeline, which includes OCR and pydicom processing of DICOM files, integration of eye gaze data (frame ID, (x, y) coordinates, duration) to create the final 3D gaze data. We create the final radiology report and clinical findings from the Report Audio.

diagnosis (classification and report generation), as shown in Fig. 1b. Especially, the scanpath prediction task on CT data is often overlooked due to missing gaze data.

The volumetric nature of CT data reveals several unique behaviors from radiologists' eye movement. For example, radiologists must constantly navigate across multiple slices and viewing planes to understand anatomical relationships [2, 22]. Additionally, radiologists mentally integrate spatial information across different depths to form a cohesive 3D understanding of the anatomy and pathology [2, 21, 24]. Finally, they employ systematic search strategies to ensure thorough volume examination, as overlooking even a single slice risks missing critical findings [2, 22, 68]. The existing scanpath prediction methods [13, 50, 80] are primarily designed for 2D imaging analysis. Consequently, these 2D-based models do not account for the unique aspects of CT interpretation, particularly the complex volumetric eye movement strategies and comprehensive slice coverage patterns exhibited by radiologists, such as moving back-and-forth between slices. Motivated by these challenges, this paper introduces the first public eye gaze medical dataset that focuses on CT scans, CT-ScanGaze. Unlike existing medical eye gaze datasets [6, 38], CT-ScanGaze provides 3D eye gaze data associated with every CT volume. As shown in Fig. 1c, CT-ScanGaze provides four main modalities: CT scans, eye gaze data (gaze map and attention map), radiology reports, and findings.

To conduct a benchmark on the proposed CT-ScanGaze, we tackle the scanpath prediction task. While existing scanpath prediction methods can be extended to 3D, their original 2D design may limit their ability to model inter-slice navigation and spatial-temporal continuity. Moreover, scaling these methods to 3D introduces a higher-dimensional search space, making them more susceptible to the curse of dimensionality and harder to generalize without specialized architectural design. Therefore, we propose a transformer-based network, CT-Searcher, that generates 3D scanpaths for CT volumes. DL typically requires large pretraining datasets, but CT-ScanGaze is relatively small with only 909 CT volumes compared to COCO-Search18's 6,202 images, potentially leading to overfitting or suboptimal training. So, we utilize a 2D-to-3D pipeline to create a synthetic 3D gaze dataset for the pretraining step. By pretraining CT-Searcher, it gains the ability to process CT features and predict gaze-like sequences, which enhances its performance on the final scanpath prediction task. Next, we train and evaluate CT-Searcher against current state-of-the-art scanpath prediction models [13, 50, 80], which are adapted to process 3D inputs and predict 3D scanpaths. Our findings indicate that CT-Searcher successfully generates radiologist-like scanpaths, effectively capturing both spatial coverage within each CT slice and navigational movements across slices.

Our contributions are summarized as follows:

- **CT-ScanGaze:** We present the first public dataset of expert radiologist gaze during CT analysis, comprising CT scans, eye gaze data, detailed reports, and findings.
- **CT-Searcher:** We introduce a 3D scanpath prediction network that generates radiologist-like eye movement from a CT volume. And a pretraining pipeline on synthetic 3D gaze data (CT, 3D gaze) from 2D gaze data (CXR, 2D gaze).

2. CT-ScanGaze

2.1. Data Collection

Our study collects data from a private hospital dataset of chest and abdomen CT scans, working with two experienced radiologists (10+ years experience). Data collection setup involves a Tobii eye tracker mounted on a monitor to track gaze data and a microphone to record audio reports. Calibration is performed before each session, with Tobii Pro Lab software managing the eye-tracking data and screen recordings. CT scans are viewed using OHIF viewer integrated with OHIF-Orthanc server, providing necessary

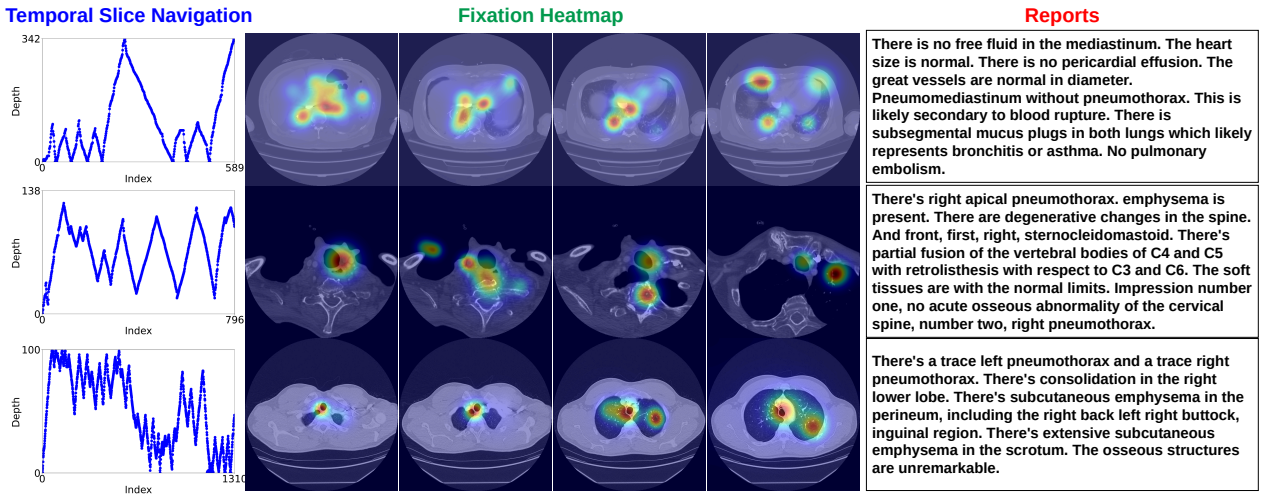


Figure 3. Examples from our dataset. Three CTs are reviewed and concluded with radiology reports as shown in the **Reports** column on the right. The **Temporal Slice Navigation** shows how slice navigation change over time. We can observe that radiologists often scan through all slices and go back-and-forth for suspicious areas. In the middle column **Fixation Heatmap**, we show some slices in the CTs. The heatmap represents all fixations viewed on that slice. For example, if in the first view they only see the left side, and in the second view they see the right side, the visualized heatmap shows fixations on both sides.

tools, e.g. contrast window control, for radiological analysis. We organize the setup as in Fig. 2 (left). Our radiologists perform their natural reading process of every CT scan exactly like they do in clinical practice. For example, 5mm images are sufficient for many pathologies. It is impractical for radiologists to review too many images in the thin series. Thin series are used for troubleshooting when radiologists need to examine certain details. Therefore, our CT scans vary in slice thickness (1-5 mm) to match this practice.

2.2. Data Processing

Each data collection session yields four data streams: video recording of radiologists' CT reading session, DICOM files of the CT scan, audio recording report, and eye gaze data. As shown in Fig. 2 (right), we process these streams to create through three main steps: Spatial Mapping, Spatiotemporal Mapping, Radiological Report & Findings Extraction. Fig. 3 shows 3 examples being viewed by our radiologists.

Spatial Mapping. We extract individual frames from session videos and use Tesseract OCR [64] to identify series and slice numbers from the bottom-left text, with manual correction when needed. This creates frame-to-DICOM mapping pairs (Fig. 2).

Spatiotemporal Mapping. The eye tracker records time (ms), screen coordinates (x, y), and fixation duration (t). Each timestamp is mapped to a frame ID using the video's frame rate (25 FPS). Using frame-to-DICOM pairs, we map gaze data to DICOM files to obtain (DICOM ID, (x, y, t)) pairs. Since each DICOM represents a CT slice number, we transform fixations into a list of 4-tuples (x, y, z, t), where z is a slice number.

Radiological Report & Findings Extraction. We use Google's Speech-to-Text 'medical dictation' model [14] on the recorded audio to generate a textual report of the radiologist' verbal interpretation for each CT scan. From these reports, we use SARLE [20], a specialized labeler for extracting findings from CT reports, to extract the corresponding radiological findings.

2.3. Dataset Statistics

CT-ScanGaze contains 909 CT scans, each accompanied by: scanpath, a radiology report, and findings. We have a total of 131,618 CT slices, 4,772 minutes of scanpath data, and 9,332 extracted findings. Due to the original gaze data containing dense and complex scanpaths with sequences averaging 543 fixations (and reaching up to 2,708 fixations per CT), we employ a simplification algorithm from the MultiMatch toolbox [18] to make the data more manageable while preserving essential gaze patterns. This process reduces sequences to an average of 222 fixations with a maximum of 1,507 fixations. Both original and simplified versions will be made available. We recommend the reader to see Appendix A for more statistical details and Appendix H for a discussion on the gaze simplification algorithm.

2.4. Scientific Benefits

This dataset represents a valuable resource for the medical imaging and computer vision communities:

- **Benchmark for 3D Scanpath Prediction:** CT-ScanGaze serves as a benchmark for predicting visual search patterns in 3D medical volumes, addressing difficulties

unique to volumetric imaging that existing 2D-focused datasets do not cover.

- **Advancement of Explainable AI**: By linking radiologists' gaze with radiology diagnosis, CT-ScanGaze supports research in explainable report generation and classification, enhancing insights into the connection between visual attention and diagnostic reasoning. This research area is actively explored in 2D medical imaging analysis by various researchers [6, 32, 38, 57, 71].

In addition, we believe CT-ScanGaze can be used in other scenarios such as study of radiologists' gaze behavior [22], advancement of general 3D scanpath prediction, radiology training [32, 38, 71], and collaborative CAD [39, 51, 53].

3. Problem Statement: 3D Scanpath Prediction

In this work, we address the 3D scanpath prediction task on CT scans. Given a CT volume $V \in \mathbb{R}^{H\times W\times D}$, our goal is to predict a sequence of N fixations $\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_N\}$. Each predicted fixation $\hat{p}_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i)$ should minimize its deviation from the ground truth fixation sequence $\{p_1, p_2, \ldots, p_N\}$, where each $p_i = (x_i, y_i, z_i, t_i)$ represents the 3D spatial location (x_i, y_i, z_i) and duration t_i of a radiologist's gaze point. Our model aims to capture both the spatial patterns and temporal dynamics of expert visual search behavior in volumetric medical imaging.

4. CT-Searcher

The overall architecture of CT-Searcher is depicted in Fig. 4. First, a Feature Extraction module takes a CT volume V and produces suitable 3D-aware representations (Sec. 4.1). Then, a Transformer Decoder uses a set of learnable queries to attend on the 3D-aware presentation and creates appropriate decoded features for decoding our desired outputs (Sec. 4.2). Finally, the latent features go through a Spatial Prediction module to produce 3D probability maps (Sec. 4.3) and a Duration Prediction module to produce durations (Sec. 4.4). All modules in CT-Searcher are trained jointly with losses defined in Sec. 4.5.

4.1. Feature Extraction

Given an input CT volume $V \in \mathbb{R}^{H\times W\times D}$, we first extract visual features $F \in \mathbb{R}^{H'\times W'\times D'\times C}$ using a visual encoder, where H', W', and D' are the reduced spatial dimensions and C is the feature dimension. These features are then projected into a compatible representation $Z = \text{MLP}(cc(F, \tau)) \in \mathbb{R}^{L\times D_m}$ for the Transformer Encoder by a Multi-Layer Perceptron (MLP), where $cc(\cdot,\cdot)$ is a concatenate operation, τ is a special learnable token representing 'stop' fixation, D_m is the transformer hidden dimension and $L = H' * W' * D' + 1$. To incorporate spatial information, we adapt the 2D sinusoidal positional encoding [8] to a 3D version. For each dimension (x, y, z) in a 3D volume, we generate position embeddings using different frequency bands:

$$\omega_k = \exp\left(-k\frac{\log(T)}{d/2}\right), \quad k = 0, \ldots, d/2 - 1 \tag{1}$$

where $T = 10000$ is the temperature parameter and $d = D_m/3$ is the embedding dimension per axis. For each spatial position pos along each axis, we compute:

$$PE_{axis}(pos) = (\sin(pos \cdot \omega_0), \cos(pos \cdot \omega_0), \ldots, \sin(pos \cdot \omega_{N-1}), \cos(pos \cdot \omega_{N-1})) \tag{2}$$

where N is the maximum length of fixation, $axis \in \{x, y, z\}$ and positions pos are normalized to $[0, 2\pi]$. The final encoding $PE(x, y, z)$ concatenates these components:

$$PE(x, y, z) = [PE_x(x); PE_y(y); PE_z(z)] \in \mathbb{R}^{D_m} \tag{3}$$

We apply 3D positional encoding on only the spatial tokens $H' * W' * D'$ of Z and pass Z to a Transformer Encoder that composes suitable 3D-aware representations $E(Z) \in \mathbb{R}^{L\times D_m}$ for the Transformer Decoder in the next step.

4.2. Transformer Decoder

Before going through Transformer Decoder, we apply 3D positional encoding again on $E(Z)$ to retain the 3D spatial information. Then, a Transformer Decoder $D(\cdot)$ uses N learnable gaze queries $Q \in \mathbb{R}^{N\times D_m}$ to attend to relevant features in $E(Z)$ to produce decoded features $R = D(Q, E(Z)) \in \mathbb{R}^{N\times D_m}$.

4.3. Spatial Prediction

The objective of Spatial Prediction module (SP) is to generate 3D fixation maps that represent the probabilistic distribution of 3D fixation coordinates over time and the the likelihood of sequence termination. The output of the transformer's decoder $R \in \mathbb{R}^{N\times D_m}$ is projected into a fixation embedding by a Fully-connected (FC) layer, which is then convolved with the encoded feature map $E(Z)$ and a softmax layer to get the 3D spatial distribution for all fixations $\hat{Y} \in [0, 1]^{N\times L}$:

$$\hat{Y} = \text{softmax}(\text{FC}(R) \otimes E(Z)^\top), \tag{4}$$

where $\otimes$ denotes the matrix multiplication.

4.4. Duration Prediction

The objective of Duration Prediction module (DP) is to generate fixation durations that reflect the probabilistic distribution of fixation durations over time. Here, we use the re-parameterization trick [19] on the encoded features R to regress them into mean values μ_t and log-variances λ_t:

$$\mu_t = \text{MLP}_{\mu_t}(R), \quad \lambda_t = \text{MLP}_{\lambda_t}(R) \tag{5}$$

$$\hat{t} = \mu_t + \epsilon_t \cdot \exp(0.5\lambda_t), \tag{6}$$

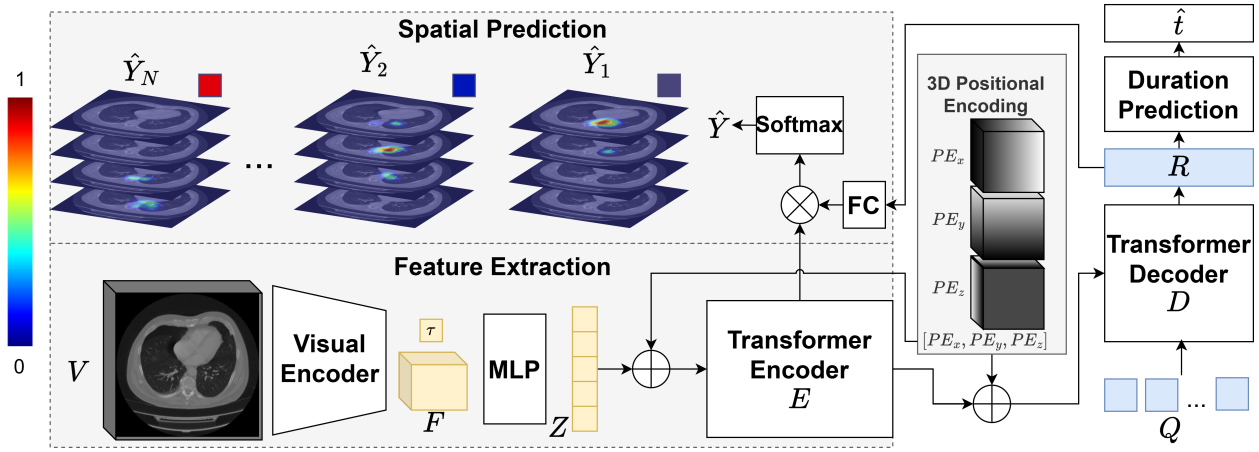


Figure 4. CT-Searcher processes CT scans V to predict 3D scanpaths. Initially, a 3D visual encoder within the Feature Extraction module extracts voxel features F. These features with a special 'stop' token are then transformed into Z by an MLP and combined with 3D positional encoding PE to incorporate spatial information. The Transformer Encoder subsequently composes suitable 3D-aware representations from Z. A Transformer Decoder then uses a set of learnable queries Q to attend to these 3D-aware representations and generate decoded features R for the desired outputs. To prevent the loss of 3D spatial information, we reapply the 3D positional encoding before feeding the 3D-aware representations into the Transformer Decoder. The decoded features R are then processed by the Spatial Prediction to produce $\hat{Y}$, including 3D fixation maps and a 'stop' probability. R is also processed by the Duration Prediction Head to estimate fixation durations $\hat{t}$ over time. Each $\hat{Y}$ has $H*W*D+1$ elements, where $H*W*D$ tokens represent the 3D fixation map and one special element, embedded from token τ, represents 'stop'. In the final $\hat{Y}_N$ of this figure, the special token has the highest probability and is likely to be chosen in sampling, signaling that the model predicts fixation stops at the N^{th} fixation.

where $\epsilon_t \sim \mathcal{N}(0,1)$ is a noise term that give our predictions a probabilistic characteristic.

4.5. Losses

To model the probabilistic nature of human fixations, we use two complementary loss functions. For the SP, which generates 3D fixation probability maps, we use Cross Entropy loss between the predicted distribution $\hat{Y}_i$ and ground truth fixation map $Y_i \in [0,1]^{L_{gt}}$ at each step i:

$$\mathcal{L}_{ce} = \frac{-1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{gt}} Y_i(l)\log \hat{Y}_i(l) \tag{7}$$

where $L_{gt} = H*W*D+1$ are the resolution of the ground truth volumes. $\hat{Y}_i$ is interpolated to same shape $Y_i \in \mathbb{R}^{H*W*D+1}$ before computing loss. The i-th ground truth 3D map U_i is initialized as zero, with the fixation location (x,y,z) as 1:

$$U_i(x',y',z') = \begin{cases} 1 & \text{if } x'=x, y'=y, z'=z \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $x' \le W, y' \le H, z' \le D$. We use 1-hot ground-truth Y framing as classification task, and CTSearch predicts distribution $\hat{Y}$ with softmax in Eq. (4). Because N is fixed, we use padding on sequences with length less than N and set the stop token τ to be 1 and $U_i = 0$, otherwise $\tau = 0$. The ground truth heatmap is $Y_i = cc(U_i, \tau)$.

The output of DP module is a scalar, making it suitable for a regression objective. We observe that the L_1 loss performs effectively in our scenario.

$$\mathcal{L}_t = \frac{1}{N}\sum_{i=1}^{N_p} \|\hat{t}_i - t_i\|_1 \tag{9}$$

where t_i is the ground truth duration at step i. The final loss combines both terms:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_{ce} \tag{10}$$

4.6. Pre-training CT-Searcher

Due to the high complexity of 3D scanpath prediction, we find that pretraining CT-Searcher before training on CT-ScanGaze is necessary to enhance its ability to process CT features and predict fixation-like sequences.

First, we combine the eye gaze data from EGD [38] and REFLACX [6], which brings more than 3,000 pairs of CXR and fixations. We also remove invalid fixation data described by the authors [6, 38] of EGD and REFLACX.

Secondly, we use a CXR-to-CT method [40] to convert all CXRs to CTs. To transform 2D gaze points into 3D coordinates, we process each normalized 2D fixation set $\{(x_i, y_i, t_i)\}_{i=1}^{n}$, where $x_i, y_i \in [0,1]$ represent normalized coordinates in the CXR image plane (with $(0,0)$ at the top-left corner and $(1,1)$ at the bottom-right corner), and t_i is the fixation duration. We perform the following steps: flip

the x dimension using $1 - x_i$ to account for the right-to-left reading pattern in CXR; set the middle slice position as 0.5 (where 0 represents the first slice and 1 represents the last slice in normalized coordinates); and map y_i directly to the slice dimension while preserving the normalized scale. The middle slice position (0.5) is chosen because radiologists exhibit a center bias when reading axial slices. This process transforms the 2D fixation set $\{(x_i, y_i, t_i)\}_{i=1}^{n}$ into a 3D fixation set $\{(1-x_i, 0.5, y_i, t_i)\}_{i=1}^{n}$ with approximately 3,000 samples.

Finally, we augment the data during pretraining by sampling fixations from a Gaussian distribution with σ of one degree of visual angle, which fluctuates coordinates in the first and second dimension. This follows the widely accepted assumption [43] that fixations follow a Gaussian distribution. We emphasize that this conversion process should only be used during pretraining due to the inherent domain gap between 2D and 3D imaging. We provide additional synthetic visualization in Appendix B.

5. Experiments

5.1. Experimental Details

Implementation Details. CT-Searcher uses Swin UNETR [29] encoder with $96 \times 96 \times 96$ input window. Due to computational constraints, we use a frozen pre-trained Swin UNETR checkpoint on LIDC-IDRI [3]. The transformer has 6 encoder/decoder layers, 8 attention heads, and hidden size of 768. All MLPs in CT-Searcher have two layers, ReLU activation, and 512 hidden dimension. The CT-Searcher is implemented in PyTorch. We use AdamW [46] optimizer with initial learning rate of $1e^{-4}$ and weight decay of 0.01. We train and evaluate all methods on the simplified version (Sec. 2.3) of CT-ScanGaze. We pre-train CT-Searcher (Sec. 4.6) for 50,000 iterations with a batch size of 8, and we then fine-tune CT-Searcher on CT-ScanGaze for 20 epochs with $N = 400$ fixations. For more implementation details, please refer to Appendix C.

Evaluation Metrics. We benchmark on two aspects: **scanpath-based** metrics (SM [16], MM [18], SED [7, 27]) and **spatial-based** metrics (CC, KLDiv, NSS [13, 30]). All metrics are adapted for 3D scanpath prediction (Appendix D). We run 5-fold cross-validation and report the scores with 95% confidence intervals.

Baselines. We evaluate CT-Searcher against leading scanpath prediction methods including PathGAN [4] Gazeformer [50], GazeformerISP [13] and HAT [80]. To benchmark these 2D methods on 3D gaze data, we replace the feature encoder of PathGAN, Gazeformer, GazeformerISP, and HAT to Swin UNETR encoder. We replace 2D positional encoding in Gazeformer, GazeformerISP, and HAT to our 3D positional encoding. Because Gazeformer has two heads for predicting the 2D coordinates and one head for predicting the duration, we only add a head to predict the slice number. In HAT and GazeformerISP, we alter their spatial decoder from generating 2D fixation heatmaps to 3D fixation heatmaps. For remaining modules, we follow their original implementation, for example we use RoBERTa [45] for task embedding of Gazeformer and GazeformerISP. More implementation details on 3D adapted scanpath prediction baselines are in Appendix E.

5.2. Qualitative Results

We present qualitative results in Fig. 5 to demonstrate the effectiveness of our proposed method. The results reveal clear performance differences across evaluation aspects. In the temporal slice navigation plots (Fig. 5, left), CT-Searcher captures similar temporal dynamics to the ground-truth, while baseline methods (GazeformerISP and HAT) exhibit erratic, high-variance patterns. Notably, the **orange areas** show that CT-Searcher replicates the non-linear behavior characteristic of experienced readers, who frequently navigate back and forth through the volume to revisit suspicious regions. The fixation heatmap visualization (Fig. 5, right) demonstrates that CT-Searcher produces more accurate attention distributions compared to other methods. These qualitative findings confirm that our approach successfully replicates the complex visual search patterns of radiologists during CT interpretation, representing a significant advancement in 3D medical scanpath prediction. We provide additional qualitative results in Appendix F.

5.3. Quantitative Results

Tabs. 1 and 2 demonstrates the significant challenges in 3D scanpath prediction. CT-Searcher achieves consistently better performance across all metrics compared to existing approaches. The earliest method, PathGAN, performs poorly across all metrics, with particularly low ScanMatch (0.0118) and saliency scores (CC: 0.0349). Recent transformer-based approaches show improvement but face various limitations: HAT struggles with limited training data (ScanMatch: 0.0171), Gazeformer's separate coordinate prediction via three MLPs causes consistency issues affecting saliency metrics (KLDiv: 23.332), and GazeformerISP underperforms on CT-ScanGaze due to its specialized and complex architecture. CT-Searcher demonstrates superior performance with significant improvements in ScanMatch (0.1466), MultiMatch position (0.7859) and duration (0.5003) scores, and saliency metrics (CC: 0.1706, KLDiv: 3.645), indicating better capture of radiologists' gaze patterns. These results highlight the effectiveness of our approach and represent a substantial advancement in modeling expert visual search behavior for CT.

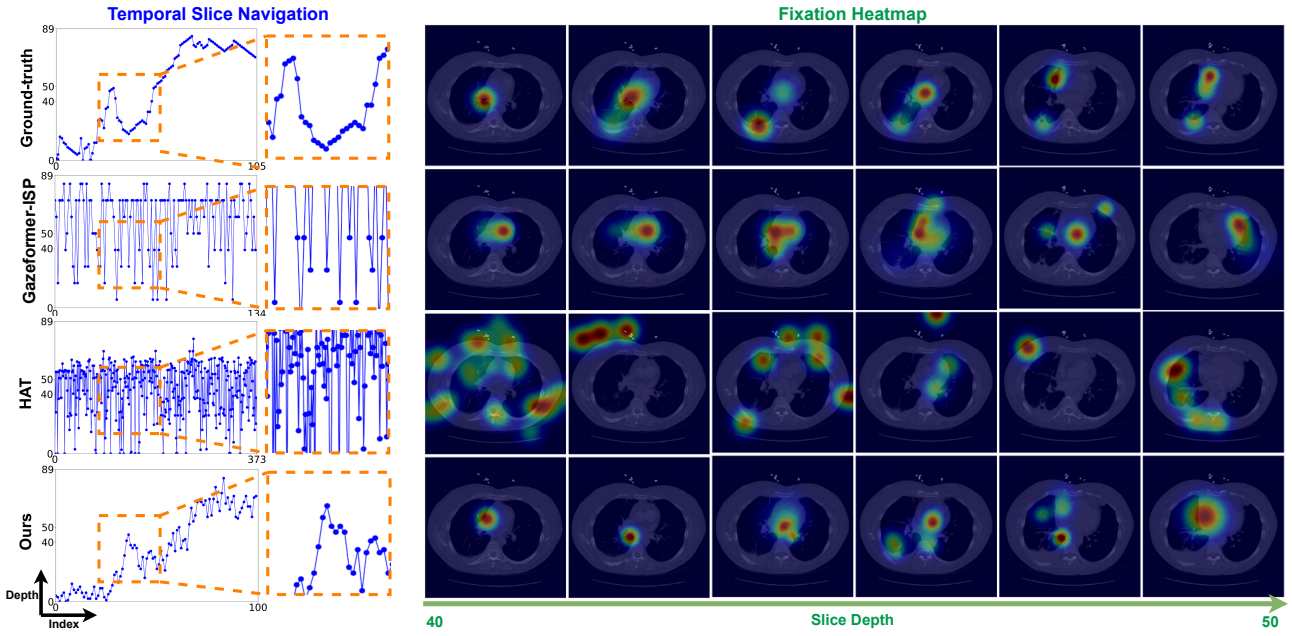


Figure 5. Qualitative comparison of our method and other SOTA methods, HAT and GazeformerISP. The **Temporal Slice Navigation** column shows the predicted scanpath over time, with the y-axis representing slice numbers (out of 89 total) and the x-axis showing fixation indices. For example, the ground truth has 105 fixations, with fixations move through depth from the earliest to the last slices. The **Fixation Heatmap** columns present the fixation heatmaps, illustrating regions of interest from which the corresponding gaze positions are sampled. The visualized CT slices for these heatmaps are selected between slice numbers 40 and 50. We observe that CT-Searcher can capture the navigation pattern of radiologists. Especially, the **orange area** indicates that the back-and-forth scanning behavior of radiologists has been successfully captured by CT-Searcher.

Table 1. Comparison of scanpath-based metrics between our CT-Searcher and existing models adapted to 3D data. (Appendix E).

Method	ScanMatch ↑		MultiMatch ↑					SED ↓
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	
PathGAN [4]	0.0118±0.002	0.0649±0.005	0.8277±0.023	0.3194±0.015	0.6786±0.031	0.6559±0.028	0.2959±0.018	663±42
HAT [80]	0.0171±0.003	-	0.8103±0.019	0.3178±0.014	0.6522±0.029	0.6295±0.025	-	307±28
Gazeformer [50]	0.0619±0.005	0.0718±0.006	0.8653±0.021	0.3012±0.016	0.8601±0.035	0.6492±0.027	0.3254±0.019	279±25
GazeformerISP [13]	0.0828±0.007	0.0711±0.006	0.8831±0.024	0.3060±0.015	0.8044±0.033	0.7354±0.031	0.3375±0.020	238±21
Ours CT-Searcher	**0.1466±0.009**	**0.1170±0.008**	**0.9216±0.026**	**0.4151±0.018**	**0.8783±0.036**	**0.7859±0.033**	**0.5003±0.024**	**174±18**

5.4. Cross-radiologist Evaluation

To assess potential learning bias toward a single radiologist's style (as CT-ScanGaze is collected from only two radiologists), we conduct a cross-radiologist evaluation. Using the same trained checkpoints from our 5-fold cross validation that produce the results in Tabs. 1 and 2, we compute separate scores for test sets containing only the first or second radiologist's ground truth data, as shown in Tabs. 3 and 4. The differences in scores between the two radiologists are insignificant, leading us to conclude that CT-Searcher successfully learns general scanpath patterns rather than overfitting to an individual radiologist's style.

Table 2. Comparison of spatial-based metrics between different scanpath prediction models.

Method	Saliency		
	CC ↑	KLDiv ↓	NSS ↑
PathGAN [4]	0.0349±0.017	25.602±1.832	0.1343±0.021
HAT [80]	0.0914±0.034	15.493±1.245	1.0676±0.089
Gazeformer [50]	0.0855±0.044	23.332±1.756	0.5005±0.042
GazeformerISP [13]	0.1104±0.021	5.023±0.428	0.7703±0.065
Ours CT-Searcher	**0.1706±0.016**	**3.645±0.312**	**1.1422±0.095**

5.5. Ablation Studies

We conduct comprehensive ablation studies to validate each component of our framework. We provide additional ablation study on CT Visual Encoder backbone in Appendix G.

Impact of 3D Positional Encoding (3D-PE). 3D-PE enhances the model's ability to capture positional relationships across height, width, and depth. As shown in Tab. 5,

Table 3. Performance of our proposed CT-Searcher on different radiologists on scanpath-based metrics.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	
Radiologist #1	0.1499±0.010	0.1214±0.009	0.9087±0.025	0.4023±0.019	0.8888±0.035	0.8001±0.034	0.5110±0.023	182±19
Radiologist #2	0.1431±0.008	0.1124±0.007	0.9339±0.027	0.4273±0.017	0.8673±0.037	0.7712±0.032	0.4891±0.025	166±17
Both	0.1466±0.009	0.1170±0.008	0.9216±0.026	0.4151±0.018	0.8783±0.036	0.7859±0.033	0.5003±0.024	174±18

applying 3D-PE improves our model performance across all metrics. Comparing rows #1 and #2 reveals substantial improvements when adding 3D-PE without pretraining, while comparing rows #3 and #4 demonstrates that 3D-PE remains crucial even with pretraining in place. The dramatic improvement in KLDiv from 7.526 to 3.645 between rows #3 and #4 highlights how 3D-PE enhances spatial awareness in our volumetric predictions. In conclusion, 3D-PE provides essential spatial context for volumetric data processing, resulting in substantial performance improvements across all metrics regardless of whether pretraining is used.

Impact of Pre-training Step (Pre). Pretraining often benefits deep learning models by establishing helpful inductive biases [76]. Given the limited size of our real CT dataset, direct training alone may lead to suboptimal performance. As shown in Tab. 5, our experiments demonstrate that incorporating a pretraining step further improves model performance. When comparing rows #2 and #4, we observe that adding pretraining to a model with 3D-PE further enhances performance, with improvements in mSM (0.1275 to 0.1318), mMM (0.6883 to 0.7002), SED (184 to 174), and KLDiv (5.194 to 3.645). These gains appear across all evaluation metrics, confirming the value of pretraining.

6. Related Work

Eye Gaze Datasets. The proliferation of eye gaze datasets in the general visual domain [10, 23, 25, 28, 36, 52, 70, 75, 77, 81] reflects growing interest in understanding human visual behavior. These datasets span diverse scenarios, ranging from multi-target search tasks [28] to focused single-category search [25, 81]. Notable examples include COCO-Search18 [77] with its extensive object categories and datasets incorporating Visual Question Answering paradigms [10]. While the general domain has seen substantial progress, medical eye gaze datasets remain limited in scope. Current medical datasets concentrate primarily on 2D modalities, particularly chest X-rays, as exemplified by EGD [38] and REFLACX [6]. The absence of 3D medical eye gaze datasets represents a significant gap in the field. To address this limitation, we present the first comprehensive eye gaze dataset for CT scan analysis. CT-ScanGaze provides essential data for advancing research in volumetric medical image analysis and understanding expert visual search patterns in 3D medical contexts.

Scanpath Prediction. Early approaches to scanpath prediction primarily focused on sampling fixations from saliency maps [34, 49, 72, 74]. The field has since witnessed remarkable progress [1, 5, 9, 11–13, 15, 17, 26, 31, 35, 41, 50, 56, 58–60, 65, 66, 73, 77–80, 82], particularly through the integration of deep neural networks [12, 17, 37, 42, 50, 60, 66, 77–79], reinforcement learning [12, 77, 78], and transformer architectures [13, 50, 60, 80]. These advances have substantially enhanced our understanding of temporal attention dynamics. However, none of previous approaches have been designed for Computed Tomography. To address this limitation, we propose a transformer-based method, CT-Searcher, specifically designed for scanpath prediction on CTs.

Table 4. Performance of CT-Searcher on different radiologists on spatial-based metrics.

Method	Saliency		
	CC ↑	KLDiv ↓	NSS ↑
Radiologist #1	0.1901±0.017	3.388±0.293	1.209±0.101
Radiologist #2	0.1503±0.014	3.912±0.331	1.072±0.089
Both	0.1706±0.016	3.645±0.312	1.142±0.095

Table 5. Ablation study on the impact of each component. **mSM** and **mMM** denote for ScanMatch and Multimatch, respectively.

Pre	3D-PE	mSM ↑	mMM ↑	SED ↓	KLDiv↓
✗	✗	0.0207±0.011	0.6034±0.031	252±23	20.958±0.902
✗	✓	0.1275±0.005	0.6883±0.022	184±18	5.194±0.523
✓	✗	0.0632±0.019	0.6562±0.019	189±12	7.526±0.616
✓	✓	**0.1318±0.009**	**0.7002±0.027**	**174±18**	**3.645±0.312**

7. Conclusion

This paper introduces CT-ScanGaze, the first public CT eye gaze dataset, along with CT-Searcher as the first CT scanpath prediction baseline. Experiments show that CT-Searcher works well and marks an important breakthrough in how we can model the way experts visually search through complex 3D medical images.

Acknowledgments. This material is based upon work supported by the National Science Foundation (NSF) under Award No OIA-1946391, NSF 2223793 EFRI BRAID, National Institutes of Health (NIH) 1R01CA277739-01.

References

[1] Hossein Adeli and Gregory Zelinsky. Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *CVPR Workshops*, 2018. 8

[2] Robert Alexander, Stephen Waite, Michael A. Bruno, Elizabeth A. Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. Mandating Limits on Workload, Duty, and Speed in Radiology. *Radiology*, 304(2), 2022. 2

[3] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 6

[4] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and NoelE O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. *ECCV Workshop on Egocentric Perception, Interaction and Computing (EPIC)*, 2018. 6, 7, 18

[5] Bahar Aydemir, Ludo Hoffstetter, Tong Zhang, Mathieu Salzmann, and Sabine Susstrunk. TempSAL - uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[6] Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1): 350, 2022. 1, 2, 4, 5, 8

[7] Stephan A Brandt and Lawrence W Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9(1):27–38, 1997. 6

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[9] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia (TMM)*, 2022. 8

[10] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 8

[11] Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He. Learning from unique perspectives: User-aware saliency modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[12] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8, 20

[13] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024. 2, 6, 7, 8, 20

[14] Google Cloud. Speech-to-text v2 api. https://cloud.google.com/speech-to-text, 2024. Accessed: 2024-03-23. 3

[15] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing (IEEE TIP)*, 2018. 8

[16] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42:692–700, 2010. 6

[17] Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 8

[18] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44:1079–1100, 2012. 3, 6, 20

[19] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 4

[20] Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021. 3, 14

[21] Trafton Drew, Karla Evans, Melissa L. H. Võ, Francine L. Jacobson, and Jeremy M. Wolfe. Informatics in Radiology: What Can You See in a Single Glance and How Might This Guide Visual Search in Medical Images? *RadioGraphics*, 33 (1), 2013. 2

[22] Trafton Drew, Melissa Le-Hoa Vo, Alex Olwal, Francine Jacobson, Steven E. Seltzer, and Jeremy M. Wolfe. Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, 13(10), 2013. 1, 2, 4

[23] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *ACM Multimedia Systems Conference (MMSys)*, 2019. 8

[24] Miguel P Eckstein, Miguel A Lago, and Craig K Abbey. The role of extra-foveal processing in 3d imaging. In *Proceedings of Spie–the International Society for Optical Engineering*. NIH Public Access, 2017. 2

[25] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009. 8

[26] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. Predicting visual importance across graphic design types. In

ACM Symposium on User Interface Software and Technology, 2020. 8

[27] Tom Foulsham and Geoffrey Underwood. What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision*, 8(2):6–6, 2008. 6

[28] Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015. 8

[29] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 6, 16

[30] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 262–270, 2015. 6

[31] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 8

[32] Bulat Ibragimov and Claudia Mello-Thoms. The use of machine learning in eye tracking studies in medical imaging: A review. *IEEE journal of biomedical and health informatics*, 2024. 4

[33] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 16

[34] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 1998. 8

[35] Sen Jia and Neil D. B. Bruce. EML-NET:an expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 2020. 8

[36] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[37] Yue Jiang, Luis A. Leiva, Hamed R. Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. UEyes: Understanding visual saliency across user interface types. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023. 8

[38] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific Data*, 8(1):1–18, 2021. 1, 2, 4, 5, 8

[39] Naji Khosravan, Haydar Celik, Baris Turkbey, Elizabeth C Jones, Bradford Wood, and Ulas Bagci. A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning. *Medical image analysis*, 51:101–115, 2019. 4

[40] Daeun Kyung, Kyungmin Jo, Jaegul Choo, Joonseok Lee, and Edward Choi. Perspective projection-based 3d ct reconstruction from biplanar x-rays. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 5

[41] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 8

[42] Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision (JoV)*, 2022. 8

[43] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*, 45(1), 2013. 6

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 18

[45] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6, 19

[46] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[47] Chong Ma, Lin Zhao, Yuzhong Chen, Sheng Wang, Lei Guo, Tuo Zhang, Dinggang Shen, Xi Jiang, and Tianming Liu. Eye-Gaze-Guided Vision Transformer for Rectifying Shortcut Learning. *IEEE Transactions on Medical Imaging*, 42 (11), 2023. 1

[48] Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, Dinggang Shen, Tianming Liu, and Xiang Li. Eye-gaze Guided Multi-modal Alignment for Medical Representation Learning, 2024. 1

[49] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research (VR)*, 2015. 8

[50] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 7, 8

[51] José Neves, Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Anderson Maciel, Andrew Duchowski, Joaquim Jorge, and Catarina Moreira. Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning. *European Journal of Radiology*, page 111341, 2024. 1, 4

[52] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014. 8

[53] João Pedrosa, Guilherme Aresta, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. Lndetector: A flexible gaze characterisation collaborative platform for pulmonary nodule screening. In *XV Mediterranean Conference on Medical and Biological Engineering and Computing–MEDICON 2019: Proceedings of MEDICON 2019, September 26-28, 2019, Coimbra, Portugal*, pages 333–343. Springer, 2020. 4

[54] Peixi Peng, Wanshu Fan, Yue Shen, Wenfei Liu, Xin Yang, Qiang Zhang, Xiaopeng Wei, and Dongsheng Zhou. Eye gaze guided cross-modal alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 1

[55] Trong Thang Pham, Jacob Brecheisen, Anh Nguyen, Hien Nguyen, and Ngan Le. I-ai: A controllable & interpretable ai system for decoding radiologists' intense focus for accurate cxr diagnoses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7850–7859, 2024. 1

[56] Trong Thang Pham, Ngoc-Vuong Ho, Nhat-Tan Bui, Thinh Phan, Patel Brijesh, Donald Adjeroh, Gianfranco Doretto, Anh Nguyen, Carol C. Wu, Hien Nguyen, and Ngan Le. Fg-cxr: A radiologist-aligned gaze dataset for enhancing interpretability in chest x-ray report generation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 941–958, 2024. 8

[57] Trong-Thang Pham, Jacob Brecheisen, Carol C Wu, Hien Nguyen, Zhigang Deng, Donald Adjeroh, Gianfranco Doretto, Arabinda Choudhary, and Ngan Le. Itpctrl-ai: End-to-end interpretable and controllable artificial intelligence by modeling radiologists' intentions. *Artificial Intelligence in Medicine*, 160:103054, 2025. 4

[58] Trong-Thang Pham, Anh Nguyen, Zhigang Deng, Carol C. Wu, Hien Van Nguyen, and Ngan Le. Interpreting radiologist's intention from eye movements in chest x-ray diagnosis, 2025. 8

[59] Trong Thang Pham, Tien-Phat Nguyen, Yuki Ikebe, Akash Awasthi, Zhigang Deng, Carol C. Wu, Hien Nguyen, and Ngan Le. Gazesearch: Radiology findings search benchmark. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 96–106, 2025.

[60] Mengyu Qiu, Yi Guo, Mingguang Zhang, Jingwei Zhang, Tian Lan, and Zhilin Liu. Simulating human visual system based on vision transformer. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, 2023. 8

[61] Graciela Ramirez-Alonso, Olanda Prieto-Ordaz, Roberto López-Santillan, and Manuel Montes-Y-Gómez. Medical report generation through radiology images: an overview. *IEEE Latin America Transactions*, 20(6):986–999, 2022. 1

[62] Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. Human attention in fine-grained classification. *arXiv preprint arXiv:2111.01628*, 2021. 1

[63] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020. 16

[64] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, pages 629–633. IEEE, 2007. 3

[65] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. ScanDMM: A deep markov model of scanpath prediction for 360° images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[66] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019. 8

[67] Ashish Verma, Aupendu Kar, Krishnendu Ghosh, Sobhan Kanti Dhara, Debashis Sen, and Prabir Kumar Biswas. Artificially generated visual scanpath improves multi-label thoracic disease classification in chest x-ray images. *IEEE Transactions on Instrumentation and Measurement*, 2024. 1

[68] Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. Interpretive Error in Radiology. *American Journal of Roentgenology*, 208(4), 2017. 2

[69] Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. 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*, pages 2194–2203, 2024. 1

[70] Shuo Wang, Ming Jiang, Xavier Morin, Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 2015. 8

[71] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 41(7):1688–1698, 2022. 4

[72] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 8

[73] Zijun Wei, Hossein Adeli, Minh Hoai, Gregory Zelinsky, and Dimitris Samaras. Learned region sparsity and diversity also predict visual attention. In *NeurIPS*, 2016. 8

[74] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[75] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision (JoV)*, 2014. 8

[76] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021. 8

[77] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse

reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[78] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 8

[79] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383*, 2023. 8

[80] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6, 7, 8, 18

[81] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8

[82] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018. 8