Abstract—Learning harmful shortcuts such as spurious correlations and biases prevents deep neural networks from learning meaningful and useful representations, thus jeopardizing the generalizability and interpretability of the learned representation. The situation becomes even more serious in medical image analysis, where the clinical data are limited and scarce while the reliability, generalizability and transparency of the learned model are highly required. To rectify the harmful shortcuts in medical imaging applications, in this paper, we propose a novel eye-gaze-guided vision transformer (EG-ViT) model which infuses the visual attention from radiologists to proactively guide the vision transformer (ViT) model to focus on regions with potential pathology rather than spurious correlations. To do so, the EG-ViT model takes the masked image patches that are within the radiologists’ interest as input while has an additional residual connection to the last encoder layer to maintain the interactions of all patches. The experiments on two medical imaging datasets demonstrate that the proposed EG-ViT model can effectively rectify the harmful shortcut learning and improve the interpretability of the model. Meanwhile, infusing the experts’ domain knowledge can also improve the large-scale ViT model’s performance over all compared baseline methods with limited samples available. In general, EG-ViT takes the advantages of powerful deep neural networks while rectifies the harmful shortcut learning with human expert’s prior knowledge. This work also opens new avenues for advancing current artificial intelligence paradigms by infusing human intelligence.

Index Terms—Eye tracking, generalizability, interpretability, shortcut learning, vision transformer.

I. INTRODUCTION

Deep neural networks have been widely used and achieved remarkable successes in many fields including natural language processing, computer vision, and medical image analysis [1], etc. Recent studies suggest that deep neural networks may be prone to learn the shortcut knowledge [2] such as the spurious correlations between the background and objects in the image (e.g., cows usually stand on the grass land) rather than the intended relevant features. Recent studies [3], [4] revealed that background is a harmful shortcut which drastically impacts the deep learning model’s performance in a negative way. The harmful shortcut knowledge, on the one hand, may not be able to generalize to new domains and tasks, and thus degenerates the performance in some scenarios such as few-shot learning (FSL). On the other hand, it jeopardizes the interpretability of the model and prevents humans from validating its underlying reasoning which is crucial in many applications, e.g., disease diagnosis with medical images.

Medical image analysis is a representative scenario where the harmful shortcut learning should be rectified because the generalizability and interpretability are highly desired and required, considering the scarcity of the clinical data (e.g., MR images with pathology) and the importance of reliability and transparency in clinical applications. The literature has already reported the existence of shortcuts in medical image analysis [5], [6]. For example, in [6], convolutional neural networks (CNNs) were employed to detect pneumonia and performed well with extremely high accuracy on the chest X-rays from a group of hospitals. However, it failed to generalize...
method to infuse the domain knowledge of an expert, i.e., rather than precious annotation in avoiding harmful shortcut in medical image applications [8], [9], [10], [11], suggests integrated eye-gaze of radiologists to improve the performance practice. Some recent deep learning studies have already been validated and widely used in longstanding clinical generalizable because it reaches the professional standard and embedded in ROIs of an expert is naturally interpretable and possible by installing the eye-tracker. Such domain knowledge be highly related to potential pathology and is easily accessible. For example, the eye-gaze information can indicate the regions-of-interest (ROIs) of radiologists, which might be exploited. For example, the eye-gaze information can indicate the regions-of-interest (ROIs) of EG-ViT are denoted by white arrows.

to the X-rays from other external hospitals with much lower performance: CNNs unexpectedly learned to detect a hospital-specific metal token at the corner of scans and utilized it for disease prediction indirectly [2], [6]. To motivate the work in this paper, in Fig. 1, we also visualize four samples of harmful shortcuts learned by vision transformer (ViT) [7] model which are the medical images’ background.

To solve this problem, one possible way is to enforce the model to concentrate on task-related objects or features rather than the harmful shortcuts by using prior knowledge [3]. For example, the bounding box and voxel/pixel-level segmentation mask of medical image directly indicate the location of the lesion on which the model should focus. However, accurate manual annotation/segmentation requires experienced radiologists and the devotion of their additional time, which is costly and not easily accessible. On the other hand, radiologists read dozens of patients’ images and write diagnosis reports on average in their routine work. This means that there is a huge amount of valuable data that is not collected and fully exploited. For example, the eye-gaze information can indicate the regions-of-interest (ROIs) of radiologists, which might be highly related to potential pathology and is easily accessible by installing the eye-tracker. Such domain knowledge embedded in ROIs of an expert is naturally interpretable and generalizable because it reaches the professional standard and has been validated and widely used in longstanding clinical practice. Some recent deep learning studies have already integrated eye-gaze of radiologists to improve the performance of medical image applications [8], [9], [10], [11], suggesting the potential usage and convenience of using eye-gaze rather than precious annotation in avoiding harmful shortcut learning.

Inspired by this, we propose an intuitive and effective method to infuse the domain knowledge of an expert, i.e., eye-gaze, with the training of deep learning models for rectifying the harmful shortcut learning in medical image analysis. Specifically, based on vision transformer (ViT) [7], we introduce a novel eye-gaze-guided vision transformer (EG-ViT) model which applies an eye-gaze mask to input image patches to screen out those irrelevant to radiologist’s visual attention and guide the model to focus on patches that are highly related to potential pathology during the model training/fine-tuning. Meanwhile, a residual connection between the unmasked input and the last ViT encoder layer is intentionally added to maintain the interactions and relationships of all patches. In the testing stage, the mask operation and residual connection are removed to maintain the original structure of ViT model.

In this way, the EG-ViT model infuses the expert’s domain knowledge to enforce the model to avoid learning harmful shortcut while takes the power of data-intensive ViT model in a more effective manner. We evaluate the proposed EG-ViT on disease diagnosis with two publicly available datasets, namely, INbreast [12] and SIIM-ACR [13]. Our extensive experiments demonstrate that the proposed EG-ViT model effectively avoids the harmful shortcut learning (Fig. 1). The diagnostic accuracy is also improved, compared with CNNs (about 4%) and ViT (about 2%) baselines with limited data. The code of our EG-ViT is available on Github.

In general, the main contributions of our work are as follows:

1) We propose a novel EG-ViT model to infuse the human expert’s prior knowledge to guide the model focusing on the region with potential pathology, avoiding the harmful shortcut learning and improving models’ interpretability with much higher performance.

2) The proposed EG-ViT model only includes an additional mask operation and a residual connection compared with vanilla ViT, thus allowing the inheritance of the parameters from a pre-trained ViT model without any additional cost.

3) We introduce a novel evaluation metric for quantifying the degree of shortcuts in models and measuring the improvement in rectifying the shortcut learning, which can also be generalized to other scenarios and tasks.

II. RELATED WORKS

A. Shortcut Learning

Deep neural networks often solve the task-specific problem, e.g., image classification, by learning the shortcuts such as the correlations between cows and grass instead of the intended solution, e.g., the features from cows [2]. Recently, the shortcut in deep learning models gains increasing attention across the deep learning field from computer vision (CV) [4], [14], [15], natural language processing (NLP) [16], [17] to reinforcement learning [18]. Various methods have been devised to mitigate the negative effects of shortcuts [3], [19], [20]. For example, in [3], a framework named COSOC was proposed to tackle this shortcut problem by extracting the foreground objects in images to get rid of background-related shortcuts based

---

1 https://github.com/MoMarky/Eye-gaze-Guided-Vision-Transformer
on a contrastive learning approach. Du et al. [19] proposed a measurement for quantifying the shortcut degree, with which a shortcut mitigation framework was introduced for natural language understanding (NLU). Shen et al. [20] forced the network to learn the necessary features for all the words in the input to alleviate the shortcut learning problem in supervised Paraphrase Identification (PI).

In medical image analysis, shortcut learning not only has a negative impact on the models’ generalizability, e.g., the same type of medical images the model performance often varies greatly between images from different vendors [21], but also degenerates the interpretability and the reliability of the applications. Recently, more studies begin to scrutinise the shortcut learning in different scenarios of medical imaging applications. For example, Luo et al. [5] demonstrated that models for the localization task are less prone to shortcut learning than models for the classification task, because the data annotation for the localization task is more fine-grained. Mahapatra et al. [22] mitigated the shortcut learning in medical image classification and segmentation by introducing an interpretability-guided inductive bias loss function which is composed of the class-distinctiveness and spatial coherent loss between the attention maps. Nauta et al. [23] made the model more focusing on the lesion area by replacing some image patches of the original image.

However, the aforementioned methods still need sufficient samples (more than 10,000) and even with fine-grained annotations such as pixel-level segmentation mask. In this paper, We rectify possible shortcut learning on small-scale medical image datasets (around 1,000) by infusing the accessible coarse eye-gaze data from radiologists. To the best of our knowledge, the current methods for investigating shortcut learning have concentrated mainly on CNN and NLP models, and the issue of shortcut learning in ViT models has not received significant scrutiny. Hence, the phenomenon of shortcut learning in ViT architectures remains to be investigated.

### B. Eye Tracking in Radiology

Visual diagnosis plays a central role in radiology, and eye-tracking procedures have proven to be a valuable tool in the study of visual diagnostic processes in radiology for decades [24]. A group of early studies have found that experts can quickly locate potential lesions with a global search and use a larger functional field of view and more conceptual knowledge than novices to find abnormalities [25], [26], [27]. Ellen et al. [28] showed that experts searched normal CXR more systematically than novices. With the rise of deep learning in computer aided diagnosis (CAD), the integration of radiologists’ eye movement into deep learning models becomes more popular. For example, N. Khosravan et al. [10] combined the radiologists’ visual attention maps with contextual information to obtain the foreground and background regions and used them for lesion segmentation. Mall et al. [29] modeled the visual search behavior of radiologists and their interpretation of mammography with CNNs. Furthermore, they [30] investigated the relationship between human visual attention and CNNs in finding lesions in mammography. Recently, Karargyris et al. [8] developed a dataset with CXR, eye-gaze, and text diagnosis reports. They proposed a multi-task framework which predicted eye-gaze and diagnosed diseases at the same time. Wang et al. [9] used radiologists’ visual attention to supervise the CNN’s attention via an attention consistency module, thus improving the diagnosis performance in osteoarthritis assessment of knee X-ray images.

Despite the successes of combining the radiologists’ eye-gaze information with CNNs, how to integrate eye-gaze information with powerful ViT model to further boost its performance in medical imaging applications still needs investigations.

### C. Vision Transformer

Since ViT [7] was introduced, transformer structure has been receiving increasing attention in the computer vision community [31]. Several effective strategies have been proposed to improve model performance and efficiency in image classification, such as knowledge distillation in DeiT [32], depth-wise convolution in CaiT [33], shifted windows in Swin Transformer [34], and tree-like structure in NesT [35]. However, the data-intensive characteristic of ViT makes it challenging to adapt to the target domain quickly with limited amount of data. To address this problem, several methods with distillation approach [32], smoothing the loss landscapes at convergence [36], incorporating CNNs like CCT [37] and locality information [38] have been proposed to reduce the demand for extensive training data to a certain extent. Nonetheless, fast adaption to the target domain still requires more innovative and effective methods to further reduce the demand for training data.

In medical image analysis, for models trained on large datasets, ViT-style models have been explored in CAD tasks on the chest X-ray (CXR) images [39]. For example, Krishnan et al. [40] and Park et al. [41] utilize ViT-based models to achieve higher COVID-19 classification accuracy through CXR images. COVID-Transformer [42] and xViTCOS [43] have been proposed to improve classification accuracy and focus on diagnosis-related regions. Bhattacharya et al. [44] combined the radiologists’ eye gaze information using a transformer model based on the teacher-student approach to effectively improve the diagnostic performance of chest X-ray disease. However, the aforementioned methods still require large amount of training data. In the scenarios with limited amount of data, it is more likely to trigger shortcut learning by exploiting the redundant information such as backgrounds, which leads to serious consequences. In addition, the small scale of medical image datasets also limits the performance of ViT. In this paper, we guide the transformer model to focus on the important regions directly by introducing the radiologists’ eye-gaze as auxiliary information and demonstrate the vanilla ViT model can handle the diagnosis task even on small scale datasets.

### III. Method

The main idea of our EG-ViT model is to utilize the eye-gaze data to mask the patches out of radiologists ROIs for...
ViT model. We first illustrate the collection of eye-gaze data and generation of eye-gaze heatmap in Section III-A. Then, we introduce the generation of eye-gaze mask used in EG-ViT model in Section III-B. Finally, we elaborate the architecture of our EG-ViT model.

A. Eye-Gaze Data Collection and Heatmap

Compared with fine-grained annotations such as pixel-level segmentation masks, coarse eye-gaze data is much more accessible during the radiologists routine works. However, we are still lacking a public available eye-gaze collection systems specially designed for radiologists for the diagnosis with the minimal interruption. Inspired by this study [11], we design a collection system specifically for capturing eye-gaze data from radiologists. Specifically, we use Tobii pro Nano as the hardware platform to collect eye-gaze data, and develop a software system that support radiologists to manipulate images freely while collecting the eye-gaze, mouse, and keyboard data simultaneously. In addition, we design a structured diagnostic report writing software that allows the radiologist to diagnosis the image and write report at the same time. The source code is publicly available at Github2. More details about the collection system can be found in the project page.

After the collection of the raw eye-gaze data, we apply two pre-processing steps. The first step is to remove the noises. As a result of blinking or turning the head, the radiologist’s eye-gaze points can fall into the areas outside of the breast tissue (such as the black background in a mammogram image), which will be filtered out. The second step is to extract the effective fixation points. Eye movements mainly consist of fixation and saccade, where saccade is a rapid movement process between two gaze points and thus does not reflect the region of attention. So we filter the saccade part and only keep the fixations of radiologists (“Processed Gaze Points” in Fig 3) by using the well-established I2MC method [45]. Finally, the eye-gaze heatmap (“Gaze Heatmap” in Fig 3) of whole image is generated by smoothing the binary eye-gaze points map through a two-dimensional Gaussian kernel with the radius of 150 pixels and the sigma 25.

B. Generation of Eye-Gaze Guided Mask

1) Image Cropping Strategy: Medical images often have a large resolution, for example, the INbreast [12] dataset used in this paper consists of images with a size of 3000 × 4000 pixels. Direct utilization of images with the original resolution requires huge amount of computational resources which is infeasible in practice especially for model training. However, compressing the whole image with a lower resolution can cause the loss of details and even miss smaller lesions. In this paper, we introduce a strategy to crop the original image into image patches and use the cropped image patches as samples for training and testing. Specifically, in the training stage, we adopt random cropping to generate image patches for each image as well as the corresponding heatmap. If a cropped image contains the lesion area, we assign a corresponding label for it. Then, we balance the number of cropped images with different labels. In the testing stage, we apply a large view sliding window to crop the whole image as patches with overlap for model evaluation.

2) Eye-Gaze Guided Mask: With the cropped eye-gaze heatmap as in previous section, we can generate the eye-gaze guided mask for the corresponding cropped image during the model training. We introduce two types of eye-gaze guided masks: Focused Eye-Gaze Mask and Separated Eye-Gaze Mask. As shown in Fig. 2, the focused eye-gaze mask is defined as a rectangular binary mask centered at the pixel with the largest value in heatmap. The separated eye-gaze mask is obtained by selecting a certain percentage of pixels according to the values in heatmap in a descending order and setting the selected positions of the mask to 1. The focused mask only keeps the greatest interest of radiologists while the separated mask tends to include all sub-regions of radiologists’ interest. If the cropped image does not have eye-gaze data, mask is not used in the training process. For comparison, we also include a random mask and generate a self-supervised mask based on model’s own attention by using Grad-CAM [46] method.

C. Eye-Gaze-Guided Vision Transformer

Compared with natural images, medical images usually have a higher resolution while pathology such as lesions locates in a small region with a noisy background, which makes model prone to learning background shortcuts rather than the intended meaningful features. To avoid learning harmful shortcuts, an intuitive idea is to guide the model to focus on the regions that are potentially related to pathology based on some prior knowledge. As discussed in section II-B, the visual attention from a radiologist during the diagnosis can serve as such prior knowledge as the guidance for the model training. In this paper, we implement this idea by introducing an eye-gaze guided mask on input image patches of ViT model to screen out the background patches. The overall architecture of EG-ViT model is shown in Fig. 3. Specifically, we first

2https://github.com/MoMarky/eye-tracking-system-for-radiologists
pre-process the collected radiologists’ eye-gaze data. Then, we generate the eye-gaze heatmap and randomly crop the original image and corresponding heatmap into a smaller size for model training. After the patch embedding, we mask out the regions out of radiologists’ ROI based on the mask generated by heatmap to make the network only focus on specific regions (i.e., ROI of radiologists). Meanwhile, to maintain the information and interaction of all patches, a residual connection is introduced in the last layer of the EG-ViT model.

1) Eye-Gaze Guided Mask Operation: With the eye-gaze guided mask, we can perform a mask operation on the input patches of the ViT model. Specifically, the input cropped image can be divided into $N$ patches where $N = (H \times W)/P^2$ is the patch number, $H$ and $W$ are the height and weight of images, $P$ is the patch size. The ViT model maps the images patches $x_i^p$ ($i = 1, 2, \ldots, N$) to $D$ dimension patch embedding $z_0 \in \mathbb{R}^{(N+1) \times D}$ (contacted with a class token) with a trainable linear projection $E \in \mathbb{R}^{P^2C \times D}$ where $C$ is the number of channels of the images:

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{\text{pos}}$$  \hspace{1cm} (1)$$

where $z_0^{0,0} = x_{\text{class}} \in \mathbb{R}^N$ is the class token for classification and $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ is the learnable position embedding. Then, the embedding of the input image patch $z_0$ is masked as:

$$\tilde{z}_0 = [z_0^{0,0}; z_0^{1:N} \odot \text{mask}]$$  \hspace{1cm} (2)$$

where $\text{mask} \in \mathbb{R}^N$ is the binary eye-gaze mask detailed in Section III-B and $z_0^{1:N} = [x_p^1 E; x_p^2 E; \ldots; x_p^N E]$ is the embedding of image patches. The masked patch embedding $\tilde{z}_0$ is then input into the first layer of ViT encoder, forcing the model only exploiting the patches with potential pathology.

2) Residual Connections Preserving Global Features: For the EG-ViT model, the forward propagation of each transformer encoder layer can be written as:

$$\tilde{z}_l = MSA(LN(\tilde{z}_{l-1})) + \tilde{z}_{l-1}$$  \hspace{1cm} (3)$$

$$\tilde{z}_l = MLP(LN(\tilde{z}_l)) + \tilde{z}_l$$  \hspace{1cm} (4)$$

where $\tilde{z}_l$ is the $l$-th layer’s embedding of masked patches. $MSA$, $MLP$, and $LN$ are the multilayer self-attention, multilayer perceptron, and layer norm in each block, respectively.

However, masking some patches in the first layer results in a risk of missing useful background information and positional relationships among all patches. So we add the initial patch embedding to the last layer’s embedding through a residual connection to maintain the information from all patches and the correlations among them. Therefore, the input of the last transformer encoder layer $\tilde{z}_l^{i-1}$ ($i=0, 1, 2, \ldots, N$) can be written as:

$$\tilde{z}_l^{i-1} = \begin{cases} 
\tilde{z}_0^{i-1}, & \text{if } i = 0 \\
\tilde{z}_0^i + \tilde{z}_0^i, & \text{if mask}_i = 0 \\
\tilde{z}_{l-1}^i + z_0^i, & \text{otherwise}
\end{cases}$$  \hspace{1cm} (5)$$

where $\tilde{z}_{l-1}$ and $\tilde{z}_{l-1}^i$ are the embeddings before and after additive operations.

3) Pre-Training and Fine-Tuning Style: As mentioned in Section II-C, although the ViT model has a strong feature representation ability, it relies heavily on large amounts of training data. However, medical images are often limited and the ViT models trained from scratch on small-scale medical images datasets often perform poorly. So instead of training from scratch, we initialize the EG-ViT parameters with the weights pre-trained on ImageNet-1K [47] and fine-tune on the small-scale medical image datasets. It is noted that EG-ViT does not introduce any additional parameters to the vanilla ViT model, thus allowing our model to directly inherit the parameters of the pre-trained models. In this way, the time required for model training is greatly reduced while the performance can be also guaranteed. Notably, the computational overhead...
of the transformer model is related to the number of patches, so by adding a mask, the resources for training EG-ViT model are further reduced.

**IV. EXPERIMENTS**

In this section, we conduct detailed experiments to demonstrate the effectiveness and advantages of EG-ViT model in rectifying the shortcut learning and improving the accuracy in diseases diagnosis. We firstly introduce the datasets used in this study for training and evaluation (Section IV-A) and then propose a new metric for quantifying the degree of shortcut learning in disease diagnosis (Section IV-B). We demonstrate the performance of EG-ViT model in shortcut rectification in Section IV-D. In Section IV-E, we compare the EG-ViT model with several baselines with and without eye-gaze data in disease diagnosis. Finally, we evaluate the effects of different masks for EG-ViT model in Section IV-F.

**A. Datasets**

To evaluate the proposed EG-ViT model, we adopt two different public clinical datasets with approval: INbreast [12] and SIIM-ACR [13], [48]. The INbreast dataset [12] includes 410 full-field digital mammography images which were collected during low-dose X-ray irradiation of the breast. We invited a radiologist with 10 years of experience to diagnose the images in this dataset and collected the complete eye movement data using the aforementioned collection system. According to BI-RADS [49] assessment of masses, these images can be classified into three groups: normal (302 cases), benign (37 cases), and malignant (71 cases), respectively. As for the SIIM-ACR dataset [13], Saab et al. [48] randomly selected 1,170 images with 268 cases of Pneumothorax and collected gaze data from three experienced radiologists. More details refer to [48].

For INbreast dataset [12], we apply the following three experimental setups: 1). In the training stage, we adopt a random cropping strategy instead of sliding window in order to ensure the balance and diversity of different samples. Specifically, we first randomly split the patients into 80% and 20% as training and testing datasets. To balance the training dataset, we perform several random cropping as well as the contrast-related augmentation for each image. Finally, our training set consists of 482 normal samples, 512 benign mass samples, and 472 malignant mass samples. 2). In the testing phase, the images of the remaining 20% patients were cropped by using a sliding window as described in Section III-B.1. The windows size is set to 1024 and the stride is set as the half of windows size, i.e., 512. For the SIIM-ACR dataset [13], the size of the original images are all 1024 × 1024, so we directly use the original image as the input for the model.

**B. Evaluation Metrics**

For the evaluation of model’s performance in disease diagnosis, we report the accuracy (ACC), area under curve (AUC), and F1-score (F1) on testing dataset. For evaluating the performance in rectifying the shortcut learning, we adopt the Structure Similarity Index Measure [50] (SSIM) to assess the similarity between model’s attention and the radiologists’ attention heatmap during the testing stage. In addition, we propose a new metric for assessing the degree of shortcut learning. In this work, we consider shortcut learning as the model completing the training task with incorrect or irrelevant information, resulting in the model focusing on the wrong region. Therefore, to measure shortcut learning in the diagnostic model, we compare the location of the model’s focus with that of the actual lesion.

In the INbreast [12] and SIIM-ACR [13] datasets, each lesion is associated with a segmentation mask that indicates its specific location and boundary (’A_mask’ in Fig. 4). During the testing phase of our EG-ViT, we employed Grad-CAM [46] to obtain the attention heatmap of the model, and used it to derive the image regions that the model attended to (’A_att’ in Fig. 4). Then, we computed the ratio of the attention regions that overlapped with the lesion regions using $A_{att}$, which denotes the intersection of $A_{att}$ and $A_{mask}$. We denoted this ratio as $O_{mm}$ (’$O_{mm}$’ in Fig. 4), where the letter ‘O’ is an abbreviation for overlap, and the first ‘m’ in the superscript denotes the model’s attention while the second ‘m’ the lesion mask. A higher $O_{mm}$ value indicates that the model is attending more to the actual lesion area, thereby minimizing shortcut learning problem.

To simplify the calculations, we set two boundary values. If the overlap ratio is greater than 0.9, we consider that the model has completely focused on the lesion area, so the $O_{mm}$ value is set to 1. If the overlap ratio is less than 0.3, we consider that the ratio is not enough to support the model to diagnose disease, so the $O_{mm}$ value is set to 0. For ratios between 0.3 and 0.9, we set the value to 0.5. This approach also reduces errors in the visualization method. Therefore, after counting and averaging the metrics across all test sets, higher values reflect that the model is paying more attention to the real lesion area, which means that there is less shortcut learning problem. For comparison, we also compute the radiologist’s eye gaze heatmap as described above and finally obtain a score of 0.53 in the INbreast [12] dataset and 0.56 in the SIIM-ACR [13] dataset. Therefore, $O_{mm}^{0.53}$ and $O_{mm}^{0.56}$ represent a standard for INbreast [12] and SIIM-ACR [13], respectively.

**C. Implementation Details**

We fine-tune the model for 60 epochs based on a cosine decay learning rate scheduler with an initial learning rate of...
Fig. 5. Harmful shortcut learning rectified by eye gaze guidance. (a) Examples from the INbreast dataset. (b) Examples from the SIIM-ACR dataset. In each panel of (a) and (b), the first row is the enhanced source image, the second row is the attention map of vanilla ViT obtained using Grad-CAM, and the third row is the attention map of our EG-ViT. Each column corresponds to the same example.

10^{-4} and 8 warm-up epochs. An Adam optimizer [51] with a batch size of 64 are used for optimization in our study. The cropped images are resized to 224×224 pixels. For all models in our experiment, we adopted open-source model weights that had already been pre-trained on ImageNet [52] and performed fine-tuning on INbreast [12] and SIIM-ACR [13] datasets. It should be noted that our EG-ViT model only uses eye-gaze data in the training stage. In the testing stage, we use the vanilla ViT architecture to load the trained weights for inference. And our experiments are five-fold. All models were trained on an internal server with 10 NVIDIA GeForce RTX 1080Ti GPUs (11GB). All experiments used the PyTorch deep learning framework [53].

D. Qualitative Analysis

In this subsection, we evaluate the performance of EG-ViT model in rectifying the shortcut learning qualitatively. For a better visualization, we employ the Grad-CAM [46] to generate the model’s attention map. Grad-CAM uses gradient to calculate the attention map of the model, which does not require any changes to the model structure and thus can be easily deployed to the ViT model.

Fig. 5 and Fig. 6 show two ways of rectification by our EG-ViT model for qualitative comparison. In Fig. 5, the enhanced source images are shown in the first row. The Grad-CAM maps from fine-tuned vanilla ViT model and from EG-ViT model are demonstrated in the second and third rows, respectively. It is observed that the Vanilla ViT model may make classification decisions from regions that are unrelated to valid human tissues, such as background edge in Fig. 5(a) or text flag in Fig. 5(b), in the absence of expert domain knowledge. However, when radiologists’ visual attention is incorporated into the EG-ViT model, it focuses on disease-related areas, such as the inner mammary region in Fig. 5(a) and lung region in Fig. 5(b). Fig. 6 provides examples of how the EG-ViT model enhances its attention to be congruent with the radiologist’s attention. The left three columns show examples from the INbreast dataset [12], and the right two columns are from the SIIM-ACR dataset [13]. In the INbreast dataset [12], the regions with mass are emphasized more by the EG-ViT model compared to the Vanilla ViT model. In the SIIM-ACR dataset [13], the Vanilla ViT model focused on a small portion of the lung region, whereas the EG-ViT model focused on more useful regions. Therefore, it can be inferred from the above examples that the decision-making of our EG-ViT model is more interpretable due to its ability to focus on disease-related areas with the help of visual attention from radiologists.

In Table I, we compare the results of different models with and without eye-gaze guidance. The comparison results in terms of SSIM and $O_{mm}$ of ViT-S and EG-ViT are also consistent with our observation that the attention heatmap generated by the EG-ViT model is more similar to the radiologist’s attention, and it focuses more on the lesion area. The number of samples with differences in EG-ViT model’s attention map compared with the ViT model’s attention was manually counted, and on average, 66% of all 410 cases showed significant differences compared with ViT in the INbreast dataset. Specifically, 38% of all cases where shortcut learning was corrected and 28% were enhanced. Qualitatively speaking, our proposed EG-ViT model can effectively mitigate the shortcut
learning problem and improve model performance. Further quantitative analysis is presented in Section IV-E.

### E. Quantitative Analysis

In top half of Table I, we first compare excellent baseline models such as U-Net [54], ResNet [55], EfficientNet [56], Vision Transformer [7], and Swin Transformer [34]. All baseline models were pre-trained on ImageNet dataset [52] and fine-tuned on the two clinical datasets. ResNet-18, although inferior to ResNet-50 and ResNet-101 in common classification metrics such as Acc., AUC, and F1 scores, exhibits a higher \( O^{mm} \) metric, indicating that its shortcut learning problem is less severe and the lower performance is primarily due to the complexity and learning capability of the model. On the other hand, although ResNet-101 and variants of EfficientNet have high classification metrics, their \( O^{mm} \) values are low, indicating that the performance of the models may be largely attributed to shortcut learning. We also find that although ViT-S performs better than other baseline models in terms of classification metrics, its \( O^{mm} \) needs improvement, potentially due to a lack of visual inductive bias [7], which makes it more prone to learning shortcut features, particularly in the medical field where training data is limited. A similar situation is observed in Swin Transformer. Therefore, high classification metrics do not necessarily indicate good performance and must be examined from the perspective of shortcut learning, which underpins our approach of introducing prior information to alleviate this issue. We also compare our work with that of Mahapatra et al. [22], which uses an interpretability-guided inductive bias approach to enhance the model’s feature distinctiveness and reduce the shortcut learning problem. Our comparison results demonstrate that introducing eye-gaze information as prior information to guide the model leads to significantly improvements in model performance and a more effective reduction of the shortcut learning problem.

In the bottom half of Table I, we compare our EG-ViT model with three recent studies [8], [9], [44] that have utilized eye-gaze for medical image classification tasks. [9] used ResNet [55] as the classification backbone, with the incorporation of visual attention from eye-gaze data to enhance osteoarthritis assessment on knee X-ray images. Karargyris et al. [8] employed a U-Net structure to classify three chest diseases and produced an attention map to compare with human attention. In [44], the authors used a teacher-student approach to diagnose diseases, where the teacher model learned the radiologists’ eye-gaze patterns during diagnosis and then the student model completed the learning task by using the same structure and learning global-focal attention information from the teacher model through an attention loss. We trained these three methods for the disease diagnosis on INbreast [12] and SIIM-ACR [13] datasets, respectively. As shown in Table I, our proposed EG-ViT model outperforms the compared methods in terms of all metrics on both datasets, except for the AUC on the SIIM-ACR dataset, which is slightly lower than that of RadioTransformer [44]. It is also observed that using eye-gaze guidance can improve the performance of ViT model and mitigate the shortcut learning problem on small datasets, which is even more effective than using CNNs with inductive biases.

### F. Ablation Study

In this subsection, we first discuss the effect of different types of masks, and then we investigate the degree of mask incorporation. Finally, we discuss the effect of residual connection.

#### 1) Mask Type: The performance comparison of four different types of masks is shown in Table II. The first row of the Table II shows the results of using the Grad-CAM [46] generated by the model as a guidance mask. To verify the efficacy of the proposed eye-gaze guidance method, we compare the results of using manual annotations as prior information,
i.e., the manually annotated lesion segmentation mask from the dataset is converted into a guidance mask to guide the model training in our EG-ViT architecture. The third row presents the results using manual annotation mask. The fourth and fifth rows shows the results of using focused mask and separated mask, respectively, as mention in Section III-B. Our results indicate that the separated gaze mask outperforms the Grad-CAM mask generated by the model itself, except for a lower SSIM score on the SIIM-ACR dataset. We posit that the Grad-CAM mask only relies on the model itself and does not address the inherent shortcut learning problem, which results in an unstable and inferior performance compared to the model guided by the radiologists’ gaze information. Additionally, we observed that the focused gaze mask performs worse than the separated gaze mask. We attribute this to the advantage of the separated regions in guiding the model to learn relationships between features that are far apart in larger images, which could be influenced by the individualized reading habits of radiologists. In situations where the radiologists’ gaze points are spread out and the saccade path is long, the use of a focused gaze mask could ignore features at other locations within the radiologists’ region of interest.

2) Mask Degree: The comparison results of the model with different degrees of separated eye-gaze mask and random mask are presented in Table III. The degree of mask operation indicates the percentage of the area that is masked in the guidance mask. For example, the top 20%, 25% or 30% patches with the highest values in the eye-gaze heatmap are selected as the areas of focus by the radiologist, resulting in the masking of 80%, 75%, or 70% of the original patches, respectively. In the bottom half of Table III, it can be seen that the performance of 60%-75% of the random mask improves with increasing mask ratio. Among the five eye-gaze masks in the upper half of Table III, the best performance is also achieved at 75%. It is believed that when ancillary information is available, the model’s diagnostic performance improves as the filtered regions become more focused on important focal regions within a certain range. Actually, our results are consistent with those reported in the MAE [57]. The density of semantic information is low in images, while the proportion of redundant information is relatively high, making it difficult for models to learn useful features from a small number of samples. Introducing a guidance mask allows the model to focus more on important regions in the image, making training more efficient and improving model performance. Compared to the small ratio of mask used in NLP models like BERT [58] (15%), a larger ratio of mask is required in images to effectively reduce the impact of redundant information on the model. Therefore, our model achieved a sweet spot at a mask ratio of 75%. Increasing the mask ratio to 80% led to a smaller filtered region, resulting in the loss of some critical features and thus causing a drop in model performance. In addition, to understand why the results for the 70% eye-gaze mask are slightly decreasing, the effect of the residual connection is further investigated in Section IV-F.3.

3) Residual Connection: In this work, the global information is involved in model training using a residual connection, which enhanced the global processing capability of the model. To assess the effect of residual connection and investigate the questions posed in Section IV-F.2, we conducted an experiment by removing the residual connection from EG-ViT and carrying out the same five degrees of guidance mask. The experimental results are shown in Table IV, where it is observed that the model’s performance gradually increased
from 60% to 75% when only local regions were involved in model training. However, it decreases to 80%, which supports the conclusion in Section IV-F.2 that within a certain threshold, the more specific the prior information guiding the model, the better the model performance. Without the residual connection, the model solely focuses on the crucial regions filtered by the eye-gaze mask and lost the global information, resulting in a decline in performance. The importance of residual connections is also illustrated by the fact that model performance improves for all five degrees after integrating residual connection. Additionally, it can be deduced that the lower results from the 70% eye-gaze mask in Table III are due to the relatively smaller impact of the residual connection’s improvement. In the future, we will work out a more stable global information completion method.

V. Discussion

Shortcut learning is that the model learns unintended features during the training phase, resulting in poor performance and generalization of the model during the testing phase. And the main reason for the occurrence of shortcut learning is insufficient data [2]. In many downstream tasks, training data are often limited in scale, which leads to the inevitable challenge of shortcut learning. To better assess the degree of shortcut learning in the model, we started from the perspective of measuring the differences between the model’s attention and the location of the real lesion. Based on it, we proposed a novel \( O^{\text{nn}} \) metric described in Section IV-B. Furthermore, as described in Section II-A, our work has initiated an exploration into the challenge of shortcut learning in ViT architecture, and has mitigated the issue by introducing human prior knowledge. This highlights the importance of prior information in training models on small-scale datasets. Despite the progress made in this manuscript, how to effectively detect and mitigate the shortcut learning problems in various computational models has not been well studied, which deserves more attention and efforts of the community. It will also be an important field of our further work.

VI. Conclusion

In this paper, we proposed a novel eye-gaze-guided vision transformer (EG-ViT) to infuse human expert’s intelligence and domain knowledge into the training of deep neural networks. This EG-ViT model is designed and implemented via the combination of eye-gaze guided mask generation and mask-guided vision transformer. The experiments on the INbreast [12] and SIIM-ACR [13] datasets demonstrated that the radiologist’s visual attention can effectively guide the model to concentrate on regions with potential pathology and achieve better performance. In particular, our EG-ViT model successfully rectifies the harmful shortcut learning and effectively improves the model’s interpretability.

Overall, this work contributes a feasible solution for rectifying the harmful shortcuts in medical imaging application. It also provides a novel insight towards advancing current artificial intelligence paradigms by infusing human intelligence. Our future works include extending and evaluating the EG-ViT framework on other types of images, e.g., natural images, with eye-tracking data for few-shot learning problems and various downstream tasks.

ACKNOWLEDGMENT

The authors would like to thank Qing Zhang and Yan Zhou from the Department of Radiology, Renji Hospital, Shanghai, China, for their contribution to the eye-gaze data.

REFERENCES

[1] E. Akleman, “Deep learning,” Computer, vol. 53, no. 9, p. 17, Sep. 2020.
[2] R. Geirhos et al., “Shortcut learning in deep neural networks,” Nature Mach. Intell., vol. 2, no. 11, pp. 665–673, Nov. 2020.
[3] X. Luo et al., “Rectifying the shortcut learning of background for few-shot learning,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 13073–13085.
[4] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, “Noise or signal: The role of image backgrounds in object recognition,” in Proc. Int. Conf. Learn. Represent., 2021, pp. 1–11.
[5] L. Luo et al., “Rethinking annotation granularity for overcoming shortcut cuts in deep learning-based radiograph diagnosis: A multicenter study,” 2021, arXiv:2104.10553.
[6] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” PLOS Med., vol. 15, no. 11, Nov. 2018, Art. no. e1002683.
[7] A. Dosovitskiy et al., “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent., 2020, pp. 1–11.
[8] A. Karargyris et al., “Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development,” Sci. Data, vol. 8, no. 1, pp. 1–18, Mar. 2021.
[9] S. Wang, X. Ouyang, T. Liu, Q. Wang, and D. Shen, “Follow my eye: Using gaze to supervise computer-aided diagnosis,” IEEE Trans. Med. Imag., vol. 41, no. 7, pp. 1688–1698, Jul. 2022.
[10] N. Khosravan et al., “Gaze2segment: A pilot study for integrating eye-tracking technology into medical image segmentation,” in Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging. Cham, Switzerland: Springer, 2016, pp. 94–104.
[11] N. Khosravan, H. Celik, B. Turkbey, E. C. Jones, B. Wood, and U. Bagci, “A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning,” Med. Image Anal., vol. 51, pp. 101–115, Jan. 2019.

TABLE IV

COMPARISON OF PERFORMANCE USING DIFFERENT DEGREE OF EYE-GAZE MASK WITHOUT RESIDUAL CONNECTION. **BOLD** DENOTES THE BEST RESULT

| Mask Type | INbreast \[12\] | SIIM-ACR \[13\] |
|-----------|-----------------|-----------------|
|           | Acc. ↑ | AUC ↑ | F1 ↑ | SSIM ↑ | \(O^{\text{nn}}\) ↑ | Acc. ↑ | AUC ↑ | F1 ↑ | SSIM ↑ | \(O^{\text{nn}}\) ↑ |
| 80% Gaze Mask | 90.54 | 82.25 | 87.73 | 0.342 | 0.43 | 82.67 | 65.31 | 82.73 | 0.243 | 0.20 |
| 75% Gaze Mask | 92.16 | 88.13 | 91.97 | 0.395 | 0.50 | 84.80 | 72.67 | 84.48 | 0.264 | 0.23 |
| 70% Gaze Mask | 91.22 | 86.71 | 88.81 | 0.364 | 0.47 | 84.20 | 71.81 | 84.03 | 0.252 | 0.20 |
| 65% Gaze Mask | 90.54 | 85.44 | 87.35 | 0.357 | 0.43 | 83.60 | 67.72 | 83.36 | 0.233 | 0.17 |
| 60% Gaze Mask | 90.54 | 83.75 | 87.69 | 0.361 | 0.36 | 82.67 | 63.84 | 81.65 | 0.214 | 0.13 |
T. McCoy, E. Pavlick, and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference,” in Proc. 57th Annu. Meet. Assoc. Comput. Linguistics, 2019, pp. 3428–3448.

T. Niven and H.-Y. Kao, “Probing neural network comprehension of natural language arguments,” in Proc. 57th Annu. Meet. Assoc. Comput. Linguistics, 2019, pp. 4658–4664.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” 2016, arXiv:1606.06556.

M. Du et al., “Towards interpreting and mitigating shortcut learning behavior of NLU models,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2021, pp. 915–929.

X. Shen and W. Lam, “Towards domain-generalizable paraphrase identification by avoiding the shortcut learning,” in Proc. Conf. Recent Adv. Natural Lang. Process. Deep Learn. Natural Lang. Process. Methods Appl., 2021, pp. 1318–1325.

C. Chen et al., “Deep learning for cardiac image segmentation: A review,” Frontiers Cardiovascular Med., vol. 7, p. 25, Mar. 2020.

D. Mahapatra, A. Poellinger, and M. Reyes, “Interpretability-guided inductive bias for deep learning based medical image,” Med. Image Anal., vol. 81, Oct. 2022, Art. no. 102551.

M. Nauta, R. Walsh, A. Dubowski, and C. Seifert, “Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis,” Diagnostics, vol. 12, no. 1, p. 40, Dec. 2021.

E. A. Krupinski, “Current perspectives in medical image perception,” Attention, Perception Psychophys., vol. 72, no. 5, pp. 1205–1217, Jul. 2010.

R. G. Svensson, “A two-stage detection model applied to skilled visual search by radiologists,” Perception Psychophys., 27, no. 11, pp. 11–16, Jan. 1980.

H. L. Kundel, C. F. Nodine, E. F. Conant, and S. P. Weinstein, “Holistic component of image perception in mammogram interpretation: Gaze-tracking study,” Radiology, vol. 242, no. 2, pp. 396–402, Feb. 2007.

T. Drew, K. Evans, M. L.-H. Võ, F. L. Jacobson, and J. M. Wolfe, “Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images?” RadioGraphics, vol. 33, no. 1, pp. 263–274, Jan. 2013.

E. M. Kok and H. Jarodzka, “Before your very eyes: The value and limitations of eye tracking in medical education,” Med. Educ., vol. 51, no. 1, pp. 1–11, Jan. 2017.

S. Mall, P. C. Brennan, and C. Mello-Thoms, “Modeling visual search behavior of breast radiologists using a deep convolution neural network,” J. Med. Imag., vol. 5, Jan. 2018, Art. no. 035002.

S. Mall, E. A. Krupinski, and C. Mello-Thoms, “Missed cancer and visual search of mammograms: What feature based machine-learning can tell us that deep convolution learning cannot,” in Proc. SPIE, Mar. 2019, pp. 281–287.

K. Han et al., “A survey on vision transformer,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 87–110, Jan. 2023.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 10347–10357.

K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, “Incorporating convolution designs into visual transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 559–568.

Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 9992–10002.

Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. O. Ari, and T. Pfister, “Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding,” 2021, arXiv:2105.12723.

X. Chen, C.-J. Hsieh, and B. Gong, “When vision transformers outperform resnets without pre-training or strong data augmentations,” in Proc. Int. Conf. Learn. Represent., pp. 1–11.

A. Hasani, S. Walton, N. Shah, A. Abuduwilli, J. Li, and H. Shi, “Escaping the big data paradigm with compact transformers,” 2021, arXiv:2104.05704.

Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, “LocaliViT: Bringing locality to vision transformers,” 2021, arXiv:2104.05707.

F. Shamshad et al., “Transformers in medical imaging: A survey,” 2022, arXiv:2201.09873.

K. S. Krishnan and K. S. Krishnan, “Vision transformer based COVID-19 detection using chest X-rays,” in Proc. 6th Int. Conf. Signal Process., Comput. Control (ISPCC), Oct. 2021, pp. 644–648.

S. Park et al., “Vision transformer for COVID-19 CXR diagnosis using chest X-ray feature corpus,” 2021, arXiv:2103.07055.

D. Shome et al., “COVID-transformer: Interpretable COVID-19 detection using vision transformer for healthcare,” Int. J. Environ. Res. Public Health, vol. 18, no. 21, pp. 11086, Oct. 2021.

A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, “XVIT-COS: Explainable vision transformer based COVID-19 screening using radiology,” IEEE Trans. Eng. Health Med., vol. 10, pp. 1–10, 2022.

M. Bhattacharya, S. Jain, and P. Prasanna, “RadioTransformer: A cascaded global-focal transformer for visual attention-guided disease classification,” 2022, arXiv:2202.11781.

M. Nyström and K. Holmqvist, “An adaptive algorithm for fixation, saccade, and gissade detection in eyetracking data,” Behav. Res. Methods, vol. 42, no. 1, pp. 188–204, Feb. 2010.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.

O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

K. Saab et al., “Observational supervision for medical image classification using gaze data,” in Medical Image Computing and Computer Assisted Intervention—MICCAI, Berlin, Germany: Springer, 2021.

L. Liberman and J. H. Menell, “Breast imaging reporting and data system (BI-RADS),” Radiologic Clinics North Amer., vol. 40, no. 3, pp. 409–430, May 2002.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–12.

O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 15979–15988.

J. D. M.-W. C. Kenton and L. K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.