Eye-gaze-guided Vision Transformer for Rectifying Shortcut Learning

Chong Ma
Northwestern Polytechnical University
mc-npu@mail.nwpu.edu.cn

Lin Zhao
University of Georgia
lin.zhao@uga.edu

Yuzhong Chen
University of Electronic Science and Technology of China
chenyuzhong211@gmail.com

Lu Zhang
University of Texas at Arlington
lu.zhang2@mavs.uta.edu

Zhenxiang Xiao
University of Electronic Science and Technology of China
zhenxiang.up@gmail.com

Haixing Dai
University of Georgia
hd54134@uga.edu

David Liu
Athens Academy
david.weizhong.liu@gmail.com

Zihao Wu
University of Georgia
zw63397@uga.edu

Zhengliang Liu
University of Georgia
zl18864@uga.edu

Sheng Wang
Shanghai Jiao Tong University
wsheng@sjtu.edu.cn

Jiaxing Gao
Northwestern Polytechnical University
2020262504@mail.nwpu.edu.cn

Changhe Li
Northwestern Polytechnical University
ChangheLi@mail.nwpu.edu.cn

Xi Jiang
University of Electronic Science and Technology of China
xijiang@uestc.edu.cn

Tuo Zhang
Northwestern Polytechnical University
tuozhang@nwpu.edu.cn

Qian Wang
ShanghaiTech University
wangqian2@shanghaitech.edu.cn

Dinggang Shen
ShanghaiTech University
dgshen@shanghaitech.edu.cn

Dajiang Zhu
University of Texas at Arlington
dajiang.zhu@uta.edu

Tianming Liu
University of Georgia
tianming.liu@gmail.com

Preprint. Under review.

Abstract

Learning harmful shortcuts such as spurious correlations and biases prevents deep neural networks from learning the meaningful and useful representations, thus jeopardizing the generalizability and interpretability of the learned representation. The situation becomes even more serious in medical imaging, where the clinical data (e.g., MR images with pathology) are limited and scarce while the reliability, generalizability and transparency of the learned model are highly required. To address this problem, we propose to infuse human experts' intelligence and domain knowledge into the training of deep neural networks. The core idea is that we infuse the visual attention information from expert radiologists to proactively guide the deep model to focus on regions with potential pathology and avoid being trapped in learning harmful shortcuts. To do so, we propose a novel eye-gaze-guided vision transformer (EG-ViT) for diagnosis with limited medical image data. We mask the input image patches that are out of the radiologists' interest and add an additional residual connection in the last encoder layer of EG-ViT to maintain the correlations of all patches. The experiments on two public datasets of INbreast and SIIM-ACR demonstrate our EG-ViT model can effectively learn/transfer experts' domain knowledge and achieve much better performance than baselines. Meanwhile, it successfully rectifies the harmful shortcut learning and significantly improves the EG-ViT model's interpretability. In general, EG-ViT takes the advantages of both human expert's prior knowledge and the power of deep neural networks. This work opens new avenues for advancing current artificial intelligence paradigms by infusing human intelligence.

Keywords: ViT, Eye Tracking, Generalizability, Interpretability, Shortcut Learning

1 Introduction

Deep neural networks have been widely used and achieved remarkable successes in many fields including natural language processing, computer vision, and medical imaging [21], among others. Recent studies suggest that deep neural networks may be prone to learning the shortcut knowledge [8] such as the spurious correlations between the background and objects in the image (e.g., cows usually stand on the grass land) rather than intended relevant features. For example, some empirical works revealed that background is a harmful shortcut which drastically affects the deep learning model's performance in a negative way [27, 55]. 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 in medical imaging.

Medical imaging 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 imaging applications [26, 41, 57]. For example, in [57], 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 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 [57, 8]. To motivate the work in this paper, in Fig. 1, we also visualize four samples of harmful shortcuts learned by vision transformer (ViT) [5] 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 by using prior knowledge [27]. For example, when conducting the diagnosis on medical images, the positions of eye-gaze from radiologists can be leveraged as additional domain knowledge. The rationale is that these positions are the regions-of-interest (ROIs) from a radiologist's perspective, which might be highly related to potential pathology. The domain knowledge embedded in ROIs from 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

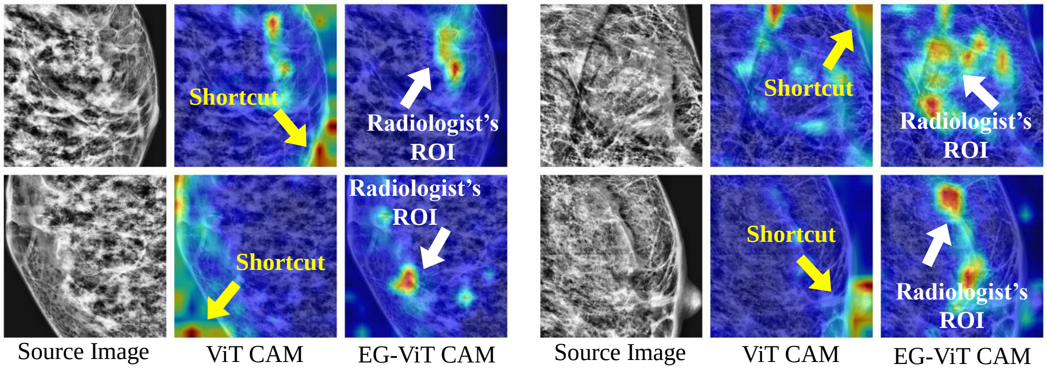


Figure 1: Illustration of the shortcuts learned by ViT model. The left columns are the source images from the public INbreast dataset. The images in the middle column correspond to the model's attention derived from Grad-CAM. It is observed that the model focuses on background shortcuts (yellow arrows) rather than the valid breast tissues. The right columns are the Grad-CAM derived from our EG-ViT model. The regions of interests (ROIs) by radiologist are denoted by white arrows.

studies have already integrated eye-gaze of radiologists to improve the performance of medical image analysis [13, 53]. However, an effective way of proactively infusing the expert's domain knowledge into deep learning processes to avoid harmful shortcut learning is still much needed.

In this paper, we propose an intuitive and effective method to infuse the domain knowledge of an expert with the training of deep learning models for FSL on disease diagnosis. Based on vision transformer (ViT) [5], 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. Meanwhile, a residual connection between the unmasked input and the last ViT encoder layer is intentionally added to retain the relationships of all patches. In this way, the EG-ViT model takes the advantages of both human expert's prior knowledge and the power of data-intensive ViT model, thus avoiding the harmful shortcut learning and infusing the expert's domain knowledge in a more effective manner. We evaluate the proposed EG-ViT on two public datasets, namely, INbreast [34] and SIIM-ACR [50]. Our extensive experiments demonstrate that the proposed EG-ViT model significantly improves the diagnosis accuracy in the FSL scenarios and successfully avoids the harmful shortcut learning compared with the baseline ViT models (Fig. 1).

In general, the main algorithmic and methodological novelties and contributions of our work are:

- We infuse the human expert's prior knowledge to guide the EG-ViT focusing on the patches with potential pathology. This design avoids the harmful shortcut learning and improves the generalizability and interpretability of the EG-ViT model with higher performance.
- The proposed EG-ViT model only introduces the mask operation and an addition residual connection, thus allowing the inheritance of the parameters from a pre-trained vanilla ViT model without any additional cost.
- Our method provides novel insights and a general framework for infusing human expert's domain knowledge into data-intensive and large-scale deep learning models such as ViT. Our work unlocks new paths for advancing current artificial intelligence paradigms by infusing human intelligence.

2 Related Works

2.1 Shortcut Learning

Deep neural networks often solve the task-specific problem, e.g., image classification, by learning the shortcuts such as the correlations of cows and grass instead of the intended solution, e.g., the features from cows [8]. Recently, the shortcut in deep learning models gains increasing attention across the deep learning field from computer vision (CV) [3, 32, 55], natural language processing (NLP) [31, 36] to reinforcement learning [1]. To date, various methods have been devised to mitigate the negative effects of shortcuts [27]. For example, in [27], 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 on a contrastive learning approach. [7] proposed a measurement for quantifying the shortcut degree, with which a shortcut mitigation framework was introduced for natural language understanding (NLU). [47] forces 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 the medical imaging field, prior works also suggested the existence of shortcuts and proposed the strategies to neutralise shortcut learning such as removing the bias in the training dataset [26, 35, 41].

2.2 Vision Transformer

Since ViT [5] was published, transformer structure has been receiving increasing attention from the computer vision community [9]. Recently, several effective strategies have been proposed to improve model performance and efficiency in image classification, such as knowledge distillation in DeiT [52], depth-wise convolution in CeiT [56], shifted windows in Swin Transformer [25], and tree-like structure in NesT [58]. However, the data-intensive characteristic of ViT makes it challenging to adapt to the target domain quickly with a limited amount of labeled data. The methods of distillation approach [52], smoothing the loss landscapes at convergence [2], and incorporating CNNs like CCT [10] and local-ViT [23] have been proposed to reduce the demand for extensive training data to a certain extent. Nonetheless, fast adaption to the target domain for FSL still requires more innovative and effective methods to reduce the demand for training data.

In the medical imaging domain, ViT-style models have been explored in computer-aided diagnosis tasks on the chest X-ray (CXR) images [46]. Krishnan et al. [17] and Park et al. [39] utilize ViT-based models to achieve higher COVID-19 classification accuracy through CXR images. COVID-Transformer [48] and xViTCOS [33] have been proposed to further improve classification accuracy and focus on diagnosis-related regions. However, there is still much room for improvement to train ViT models in a small dataset, such as medical imaging dataset.

2.3 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 [18]. Back in 1978, Nodine et al. [37] proposed a three-stage theory of visual search and detection, that is, distinguishing between initial overall pattern recognition, focused attention to image details, and final decision making. Then, the global-focal search models [51, 20, 6] further optimized the interpretation of eye-movement behavior. They also 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. Krupinski [19] reported that in mammograms with more than one lesion, experts showed more comparison scanning between the left and right breast. Kok et al. [15] found that experts' visual search patterns were more diffuse than novices. Then the authors [16] showed that experts searched normal CXR more systematically than novices. Overall, existing literature studies of eye movement in radiology and inter-individual variation provide sufficient justification to integrate this wealth of ancillary information in computer-aided diagnostic (CAD) systems in radiology.

With the rise of deep learning in CAD, the combination with eye movement is also increasing. Mall et al. [29] modeled the visual search behavior of radiologists and their interpretation of mammography using CNNs. Furthermore, they [30] investigated the relationship between human visual attention and CNNs in finding lesions in mammography. Recently, Karargyris et al. [14] developed a dataset with CXR, gaze, and text diagnosis reports. They proposed a multi-task framework which predicted gaze and diagnosed diseases at the same time. Wang et al. [53] 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.

2.4 Few-Shot Learning in Medical Imaging

FSL is proposed for learning with only a few samples [54]. Therefore, FSL is quite suitable for medical image classification due to the difficulty of obtaining a dataset with large and consistently labeled medical images. Ma et al. [28] adopts a k-Nearest-Neighbor attention pooling layer to construct the AffinityNet model, which can learn from a small number of training data and perform well in generalization. Puch et al. [40] applies an FSL model based on Triplet Network to accomplish classification of brain images. MetaCOVID [49], which is an FSL model based on Siamese Network, is proposed to classify COVID-19 cases from CXR images, and achieves a promising performance. The methods mentioned above mainly uses CNNs to extract features and they may have limited interpretability of the model. Li et al. [22] present polymorphic transformer, which exists between feature extractor and a task head and reduces the gap between different domains. This work achieves high performance on medical segmentation tasks and shows the flexibility of transformer in FSL. Although transformer-based FSL has made great progress in computer vision, more works are needed to take the full advantage of the powerful representation learning capability of ViT in medical imaging.

3 Method

The method section is organized as follows: We firstly illustrate the design of EG-ViT in Section 3.1. Then, we introduce the pre-processing of eye-gaze data and the generation of eye-gaze mask in Section 3.2.

3.1 Eye-Gaze-Guided Vision Transformer

We introduce the detailed architecture of the proposed EG-ViT in this subsection. Compared with natural images, medical images generally have a higher resolution while pathology such as lesions usually locates in a small region with a noisy background, making it difficult for the extraction of the meaningful features. To avoid learning the useless or harmful shortcuts rather than intended meaningful features, an intuitive idea is to guide the model to focus on the regions that are potentially related to pathology based on prior domain knowledge. The visual attention from a radiologist during the diagnosis can provide such domain knowledge as the guidance for model training. In this paper, we implement this idea by introducing a eye-gaze guided mask on input image patches and an additional residual connection on the EG-ViT model. The architecture of EG-ViT model is shown in Fig. 2.

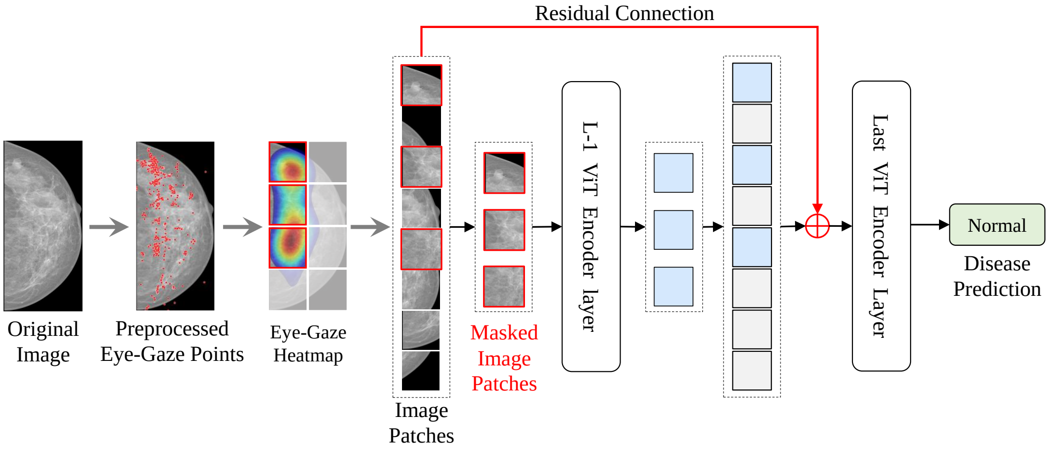


Figure 2: The architecture of our EG-ViT model. The eye-gaze points are collected and pre-processed to generate the eye-gaze heat map as masks. Then, we applied the mask to the patches derived from the original image to screen out the patches that may not be related to pathology and radiologists' interest. The masked image patches (highlighted by red rectangular) are treated as input to the transformer encoder layer. Note that to maintain the relationships among all the patches, we added an additional residual connection (highlighted by red arrow) between the input and the last encoder layer.

3.1.1 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 image can be divided into N patches where $N = (H \times W)/P^2$ is the patch numbers, H and W are the height and weight of images, P is the patch size. The ViT model maps the images patches x_p^i ($i = 1, 2, \cdots$,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 $E \in \mathbb{R}^{HWC\times D}$ where C is the channels of the images:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \cdots ; x_p^N E] + E_{pos} \tag{1}$$

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

$$\tilde{z}_0 = [z_0^0; z_0^{1:N} \odot mask] \tag{2}$$

where $mask \in \mathbb{R}^N$ is the binary eye gaze mask detailed in Section 3.2 and $z_0^{1:N} = [x_p^1 E; x_p^2 E; \cdots ; x_p^N E]$ is the image embedding patches. The masked patch embedding $\tilde{z}_0$ is then input into the first layer of ViT encoder, forcing the model focus on the corresponding patches with potential pathology.

3.1.2 Residual Connections Preserving Global Features

For vision transformer [5], the forward propagation of each transformer encoder layer can be written as:

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

$$\tilde{z}_l = MLP(LN(\tilde{z}'_l)) + \tilde{z}'_l \tag{4}$$

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

However, masking some patches in the first layer results in a risk of missing useful background information and positional relationships among blocks. Inspired by [12, 11], we add the whole initial embedding patches back to the last layer's embedding patches to retain global information and maintain the correlations of all patches. Therefore, the input embedding patch of last transformer encoder $\hat{z}_{l-1}^i$ (i=0,1,2 $\cdots$ N) can be written as:

$$\hat{z}_{l-1}^i = \begin{cases} \tilde{z}_{l-1}^0, & if\ i = 0 \\ z_0^i, & if\ mask_i = 0 \\ \tilde{z}_{l-1}^i + z_0^i, & otherwise \end{cases} \tag{5}$$

where $\hat{z}_{l-1}$ and $\tilde{z}_{l-1}$ are the after and before additive operations of the embedding patches before the last transformer encoder.

3.1.3 Pre-training and Fine-tuning Style

It should be noted that we do not add any additional parameters to the model, thus allowing our model to inherit the parameters of the pre-trained model directly without additional operation. Also, the computation of the transformer model is related to the number of patches, and by adding a mask, we also reduce the computation of ViT. Pre-training models on large datasets and fine-tuning on specific dataset is a popular paradigm for FSL. In this study, we initialize the model parameters with the model pre-trained on ImageNet-1K [43].

3.2 Generation of Eye-Gaze Guided Mask

As discussed in Section 3.1, the visual attention of radiologist plays a central role in our EG-ViT model which provides a wealth of expert's domain knowledge. Here, we introduce the steps to generate the eye-gaze guided mask used in our model.

Firstly, the eye-tracking data (i.e., the eye-gaze points) is collected when the radiologist read the medical images such as X-ray for diagnosis. More details for eye-tracking data acquisition can be found in supplemental materials. With the raw eye gaze points from the eye tracker, we filter the eye movements such as Saccades, Smooth Pursuits Nystagmus [42], and only keep Fixations of radiologists (red dots in Fig. 3) by using the well-established I2MC [38] method.

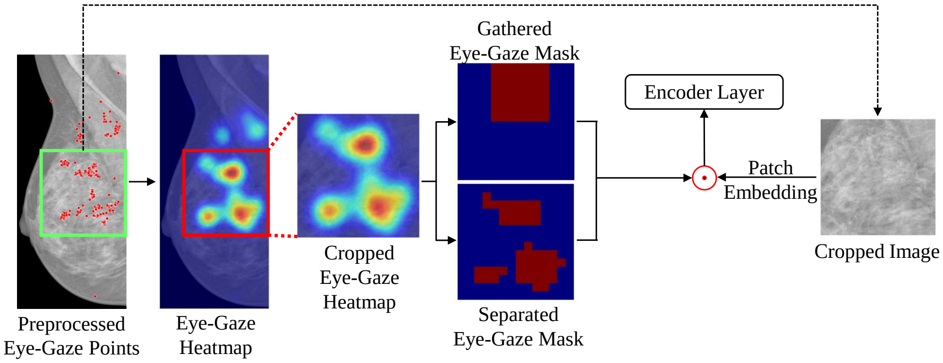


Figure 3: The generation of gathered/separated eye-gaze masks. The eye-gaze heat map is firstly cropped into the size of 224×224, based on which gathered/separated eye-gaze masks are generated, respectively. Then, these eye-gaze masks are used to mask the corresponding patch embedding, which is the input to the EG-ViT encoder layer.

Then, the eye-gaze heatmap is generated by Gaussian filtering of the eye-gaze scatter points as shown in Fig. 3. Next, to facilitate the training of the model, we crop a ROI from both the original image and eye-gaze heat maps. As the patch embedding layer maps the input image to 14×14, we also resize the eye-gaze heat map to 14×14.

Last, as shown in Fig. 3, the binary Separated Eye-Gaze Mask is obtained by setting the largest 49 values to 1 and zeroing the rest. The Gathered Eye-Gaze Mask is defined as a 7×7 binary mask centered at the largest value. The gathered mask can filter out the areas but the greatest interest to the radiologist while the separated mask tends to focus on all sub-regions of interest to the radiologist.

Notably, the combination of eye-gaze guided mask generation and mask-guided vision transformer (EG-ViT) offers an effective and powerful mechanism to infuse radiologist's domain knowledge into the representation learning of the most informative, explainable and generalizable features for medical diagnosis. Our experimental results in the next section will demonstrate that this EG-ViT model can effectively rectify harmful shortcut learning and thus significantly improve the quality of learned features with relatively small training dataset.

4 Experiments

To demonstrate the value of introducing human experts' experience and prior knowledge into deep neural network training, we apply the proposed EG-ViT model to two different public clinical datasets (Section 4.1) with eye tracking data from human radiologists. We firstly compare the prediction performance with two baselines, ResNet [12] and Vision Transformer [5] in Section 4.3. Secondly, we provide a detailed analysis of how EG-ViT can rectify shortcut in Section 4.4. Thirdly, we compare with other classification methods that used eye gaze data in Section 4.5, followed by ablation study in Section 4.6.

4.1 Datasets and Evaluation Metrics

We conduct the experiments on two datasets: INbreast [34] and SIIM-ACR [44, 50]. The INbreast dataset [34] includes 410 full-field digital mammography images which were collected during low-dose X-ray irradiation of the breast. And we invited a radiologist with 10 years of experience to diagnose the images in this database and collected the complete eye movement data using our acquisition system. According to BI-RADS [24] assessment of masses, the images can be classified into three groups: normal(302), benign(37) and malignant(71), respectively. Saab et al. [44] randomly selected 1,170 images, with 268 cases of pneumon, from the SIIM-ACR Pneumothorax dataset [50]

Table 1: The disease prediction accuracy compared with baseline ResNet and ViT model in terms of Accuracy, F1 and AUC scores. The number of parameters in each model is also reported. Red and blue denote the best and the second-best results, respectively.

Method	Params	INbreast [34]			SIIM-ACR [50]		
		Acc.	F1	AUC	Acc.	F1	AUC
ResNet-18 [12]	11M	84.34	85.55	89.04	71.67	67.85	72.42
ResNet-50 [12]	24M	91.57	91.92	94.62	69.17	67.3	69.25
ResNet-101 [12]	43M	92.07	92.55	91.52	70.83	65.33	70.31
ViT-S [5]	22M	83.13	81.06	85.36	81.20	81.62	74.46
ViT-B [5]	89M	87.95	81.35	90.86	86.00	73.69	86.67
EG-ViT (ours)	22M	92.77	92.81	94.16	85.60	84.87	74.10

and collected gaze data from three radiologists. We randomly split the dataset into 80% and 20% as training and testing dataset. For each experiment, we report the accuracy (ACC), F1-score (F1) and area under curve (AUC) on test dataset.

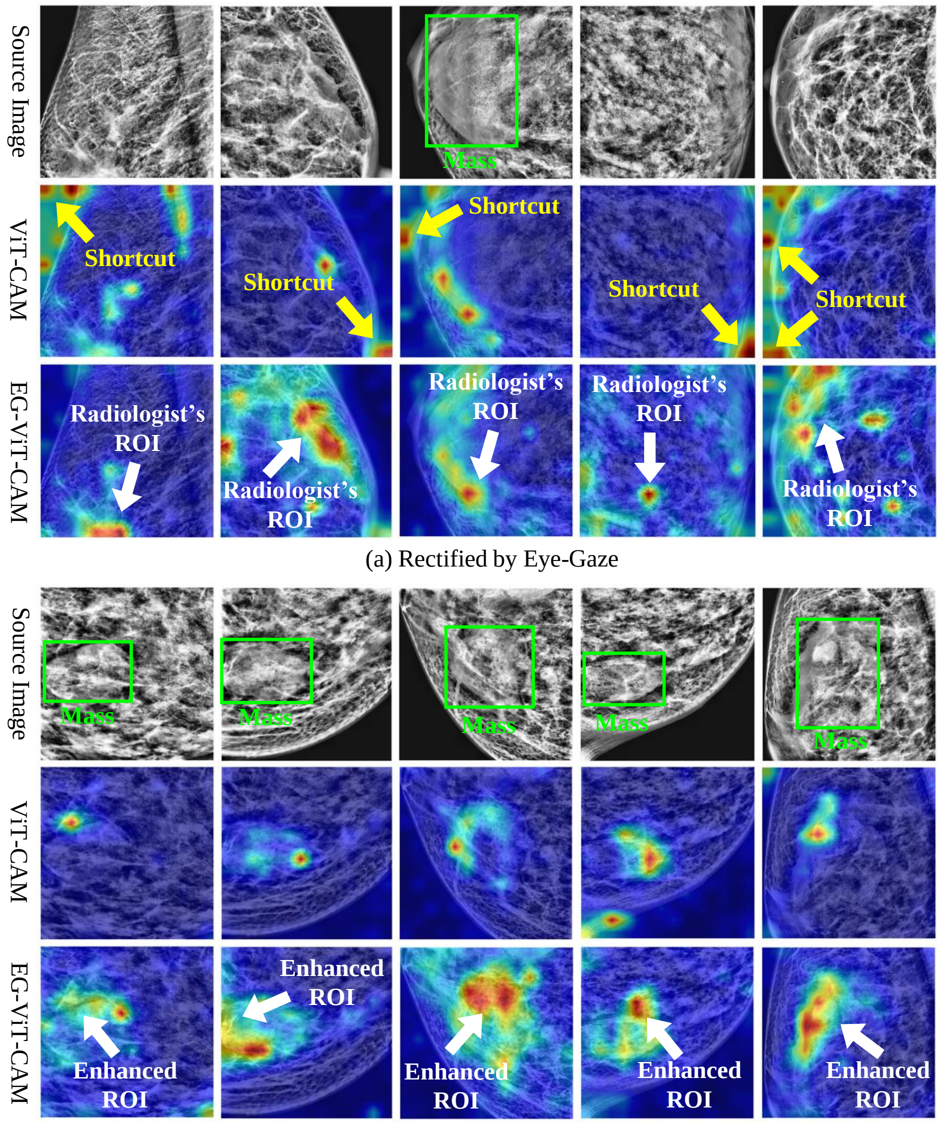
4.2 Implementation Details

We train 70 epochs with 0.0001 initial learning rate and choose a cosine decay learning rate scheduler and 8 epochs of warm-up. An Adam optimizer with a batch size of 64 are applied in our study. For the INbreast dataset [34], images and corresponding eye-gaze heatmap are with resolution of 3000×4000 pixels. For a better performance, we crop the ROI images as input images. For images with masses (include benign and malignant classes), we only crop the ROI containing the mass area as a sample. Since there are multiple masses within a single image, we count the two immediately adjacent masses as just one ROI, and finally obtain a total of 114 ROIs of mass in 108 images, of which 40 are benign and 74 are malignant. Considering the larger image field of view, we choose the size of ROI as 1024×1024. For the normal category without mass, we randomly crop only one ROI image for each image. Therefore, we obtain 302 normal, 40 benign and 74 malignant samples. The INbreast dataset [34] is divided into training and testing dataset based on samples instead of images. To balance the numbers of training samples of each category, we randomly crop the ROI 8 times for benign category and 4 times for malignant category, that is, the mass will appear at random location in the ROI images. Note that, for benign and malignant case in testing sets, the mass is centered in ROI image. Finally, the training dataset contains a total of 482 normal samples, 512 benign mass samples, and 472 malignant mass samples. We also use [59] to perform contrast enhancement twice with different threshold parameters for each cropped image in training sets. The cropped ROI images are both resized to 224×224 pixels finally. For the SIIM-ACR dataset [50], the eye-gaze heat map and X-ray images are simply resized to 224×224 pixels with no ROI cropping.

4.3 Comparison with Baselines

Here, we adopt ResNet [12] and Vision Transformer [5] as two baselines for comparison. The baseline model was pre-trained on ImageNet dataset [4] and fine-tuned on the two clinical dataset. The experimental results and parameters of each model are reported in Table 1. On the INbreast dataset [34], our EG-ViT model uses a relatively small number of parameters. Our model outperforms all the baselines in terms of F1 scores. We also compare with two baseline ViT models, ViT-S and ViT-B. Among them, ViT-S use 384 as hidden size and 12 encoder layers, ViT-B use 768 as hidden size and 12 encoder layers. It is observed that the performance of the two baseline ViT models is inferior to that of ResNets. This is because ViT has a more complex model architecture as well as a larger function space than ResNet and lacks some inductive biases that only CNNs have, making it difficult to pre-train and fine-tune on a small dataset like INbreast [34]. However, with the guidance of eye-gaze from the radiologist, the performance of ViT-based EG-ViT model is significantly improved, which suggests that eye-gaze serves as a strong prior guidance to assist network training and reduces the potential over-fitting problem caused by insufficient samples.

On the SIIM-ACR dataset [50], EG-ViT outperforms all the other methods in F1 measure, and ViT-B has higher accuracy and AUC score than the rest of the models. However, ViT-B has a much larger



(a) Rectified by Eye-Gaze

(b) Enhanced by Eye-Gaze

Figure 4: (a) Harmful shortcut learning rectified by eye gaze guidance. (b) Useful feature learning enhanced by eye gaze guidance. In each panel of (a) and (b), the first row is the source image, the second row is the attention map of ViT obtained using Grad-CAM, and the third row is the attention map of EG-ViT. Each column corresponds to the same example.

number of parameters and is more difficult to train, while EG-ViT is smaller and achieves essentially the same level of accuracy as the larger model.

4.4 Evaluation of Shortcut Rectification

In this subsection, we evaluate the performance of EG-ViT model in rectifying the shortcut learning both qualitatively and quantitatively. For better visualization of the comparison, we employ the Grad-CAM module [45] to generate the network 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.

Table 2: Comparison with other eye-gaze-guided networks. Red and blue denote the best and the second-best results, respectively.

Method	INbreast [34]			SIIM-ACR [50]		
	Acc.	F1	AUC	Acc.	F1	AUC
ResNet-18+Gaze [53]	86.74	88.32	91.48	70.0	70.07	68.94
ResNet-50+Gaze [53]	90.36	90.95	93.06	70.83	69.22	70.58
ResNet-101+Gaze [53]	91.57	91.92	94.62	72.5	71.22	72.66
U-Net+Gaze [13]	86.07	85.36	92.98	81.10	80.33	68.84
EG-ViT (ours)	92.77	92.81	94.16	85.60	84.87	74.10

Table 3: Comparison of performance using different masks. Red and blue denote the best and the second-best results, respectively.

Method	INbreast [34]			SIIM-ACR [50]		
	Acc.	F1	AUC	Acc.	F1	AUC
ViT-S [5] (Baseline)	83.13	81.06	85.36	81.20	81.62	74.46
Gathered Grad-CAM Mask	83.13	84.08	90.36	82.00	76.54	57.71
Separated Grad-CAM Mask	83.13	84.13	89.65	82.40	79.74	64.18
Gathered Eye Gaze Mask	79.52	81.45	88.89	84.00	81.58	65.9
Separated Eye Gaze Mask	92.77	92.81	94.16	85.6	84.87	74.1

Fig. 4 shows two ways of rectification by our EG-ViT model for qualitative comparison. In Fig. 4(a), the source images are shown in the first row. The Grad-CAM maps from fine-tuned vanilla ViT model and from EG-ViT are demonstrated in the second and third rows, respectively. It is observed that without the expert's domain knowledge, ViT model is likely to make classification decisions from the areas that are related to regions, such as background edge, other than valid human tissues. However, with the help of visual attention from radiologists, the EG-ViT model is guided to the meaningful areas, such as the inner mammary region. Fig. 4(b) demonstrates the cases that EG-ViT model guides and enhances the model's attention to the radiologist's ROI. The regions with mass are more emphasized by EG-ViT compared with vanilla ViT, which makes the decision of the model more interpretable. To measure the improvement of EG-ViT model quantitatively, we manually check the Grad-CAM to examine if the harmful shortcuts exist in ViT and are rectified by EG-ViT in INbreast dataset [34]. We find that EG-ViT has a significant improvement on 64% (264 cases) of all 410 cases compared with ViT. Among them, 151 cases are significantly corrected and 113 are significantly enhanced. Therefore, the radiologist's attention plays a crucial role in shortcut rectification.

4.5 Comparison with Eye-Gazed based Classification Network

In this subsection, we compare our EG-ViT model with two recent studies that also utilized eye-gaze for medical image classification tasks [53, 13] . In [53], ResNet [12] was used as the classification backbone, with which visual attention from eye-gaze data was incorporated to enhance osteoarthritis assessment on knee X-ray images. Karargyris et al. [13] used a U-Net structure to classify three chest diseases. And the model also can output attention map to compare with human attention map. We train these two methods for the disease classification task on INbreast [34] and SIIM-ACR [50], respectively. As shown in Table 2, our proposed EG-ViT model outperforms the other methods on both INbreast [34] and SIIM-ACR datasets [50] in terms of most metrics, except for AUC for INbreast [34], on which our model yields the second-best performance. It can be seen that the aid of eye-gaze heat map can improve the performance of ViT model on small datasets, even beyond the CNN models.

4.6 Ablation Study

In this subsection, we discuss the effect of the eye gaze mask. As shown in Fig. 3, we conduct comparative experiments on separated position mask and gathered position mask. The comparison of performance using these two types of masks is shown in Table 3. The first row of the table shows the results of training with the ViT-S [5]. In the second and third rows, we use the Grad-CAM generated by the network and compare the effect of gathered and separated position masks, and the fourth and fifth rows show the comparison of the two masks generated by human attention heat-map. In our experiments on both datasets, we find that the separated Grad-CAM mask is slightly better than the gathered one, especially for eye gaze mask. This could be attributed to that separated regions have the advantage of helping the model learn the relationship between features that are further apart in larger images. Also, when a radiologist has strong prior knowledge in the region out of the mask, the eye gaze mask will have a more scattered distribution, due to the flexibility of human attention itself and the different reading patterns of experts. We also see that the result of gathered gaze position masks is worse than Grad-CAM mask on the INbreast dataset [34]. This may be related to the radiologists' individualized reading habits. Specifically, if the radiologists' gaze points are spread out and the saccade path is long, then the use of a gathered position mask will ignore the features at the other locations where the radiologists focus on. We will continue to optimize the way that gaze mask is generated when we collect more eye gaze data from radiologists in the future.

5 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 [34] and SIIM-ACR [50] datasets demonstrated that radiologist's visual attention can effectively guide the EG-ViT model to concentrate on regions with potential pathology and achieve much better performance compared with baselines. In particular, our EG-ViT model successfully rectifies the harmful shortcut learning and significantly improves the model's interpretability, generalizability, and performance. Overall, this work contributes novel insights and a concrete new method for 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.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)

[2] Chen, X., Hsieh, C.J., Gong, B.: When vision transformers outperform resnets without pretraining or strong data augmentations. arXiv preprint arXiv:2106.01548 (2021)

[3] Dancette, C., Cadene, R., Teney, D., Cord, M.: Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1574–1583 (2021)

[4] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[6] Drew, T., Evans, K.K., Võ, M.L.H., Jacobson, F.L., Wolfe, J.M.: Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? Radiographics **33**, 263–274 (2013)

[7] Du, M., Manjunatha, V., Jain, R., Deshpande, R., Dernoncourt, F., Gu, J., Sun, T., Hu, X.: Towards interpreting and mitigating shortcut learning behavior of nlu models. arXiv preprint arXiv:2103.06922 (2021)

[8] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)

[9] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[10] Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H.: Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704 (2021)

[11] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021)

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

[13] Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., 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)

[14] Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M.H., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., Moradi, M.: Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. Scientific Data **8**, 92–92 (2021)

[15] Kok, E.M., de Bruin, A.B.H., Robben, S.G.F., van Merriënboer, J.J.G.: Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. Applied Cognitive Psychology **26**, 854–862 (2012)

[16] Kok, E.M., Jarodzka, H.: Before your very eyes: the value and limitations of eye tracking in medical education. Medical Education **51**, 114–122 (2017)

[17] Krishnan, K.S., Krishnan, K.S.: Vision transformer based covid-19 detection using chest x-rays. In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC). pp. 644–648. IEEE (2021)

[18] Krupinski, E.A.: Current perspectives in medical image perception. Attention Perception & Psychophysics **72**(5), 1205–1217 (2010)

[19] Krupinski, E.A.: Visual scanning patterns of radiologists searching mammograms. Academic Radiology **3**, 137–144 (1996)

[20] Kundel, H.L., Nodine, C.F., Conant, E.F., Weinstein, S.P.: Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology **242**, 396–402 (2007)

[21] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

[22] Li, S., Sui, X., Fu, J., Fu, H., Luo, X., Feng, Y., Xu, X., Liu, Y., Ting, D.S., Goh, R.S.M.: Few-shot domain adaptation with polymorphic transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 330–340. Springer (2021)

[23] Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)

[24] Liberman, L., Menell, J.H.: Breast imaging reporting and data system (bi-rads). Radiologic Clinics of North America **40**, 409–430 (2002)

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)

[26] Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., Heng, P.A.: Rethinking annotation granularity for overcoming deep shortcut learning: A retrospective study on chest radiographs. arXiv preprint arXiv:2104.10553 (2021)

[27] Luo, X., Wei, L., Wen, L., Yang, J., Xie, L., Xu, Z., Tian, Q.: Rectifying the shortcut learning of background for few-shot learning. Advances in Neural Information Processing Systems **34** (2021)

[28] Ma, T., Zhang, A.: Affinitynet: semi-supervised few-shot learning for disease type prediction. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 1069–1076 (2019)

[29] Mall, S., Brennan, P.C., Mello-Thoms, C.: Modeling visual search behavior of breast radiologists using a deep convolution neural network. Journal of medical imaging **5**, 035502–035502 (2018)

[30] Mall, S., Krupinski, E., Mello-Thoms, C.: Missed cancer and visual search of mammograms: what feature-based machine-learning can tell us that deep-convolution learning cannot. In: Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment. vol. 10952, pp. 281–287. SPIE (2019)

[31] McCoy, R.T., Pavlick, E., Linzen, T.: Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007 (2019)

[32] Minderer, M., Bachem, O., Houlsby, N., Tschannen, M.: Automatic shortcut removal for self-supervised representation learning. In: International Conference on Machine Learning. pp. 6927–6937. PMLR (2020)

[33] Mondal, A.K., Bhattacharjee, A., Singla, P., Prathosh, A.: xvitcos: Explainable vision transformer based covid-19 screening using radiography. IEEE Journal of Translational Engineering in Health and Medicine **10**, 1–10 (2021)

[34] Moreira, I., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic Radiology **19**, 236–248 (2012)

[35] Nauta, M., Walsh, R., Dubowski, A., Seifert, C.: Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. Diagnostics **12**(1), 40 (2022)

[36] Niven, T., Kao, H.Y.: Probing neural network comprehension of natural language arguments. arXiv preprint arXiv:1907.07355 (2019)

[37] Nodine, C.F., Kundel, H.L.: Using eye movements to study visual search and to improve tumor detection. Radiographics **7**, 1241–1250 (1987)

[38] Nyström, M., Holmqvist, K.: An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behavior research methods **42**(1), 188–204 (2010)

[39] Park, S., Kim, G., Oh, Y., Seo, J.B., Lee, S.M., Kim, J.H., Moon, S., Lim, J.K., Ye, J.C.: Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. arXiv preprint arXiv:2103.07055 (2021)

[40] Puch, S., Sánchez, I., Rowe, M.: Few-shot learning with deep triplet networks for brain imaging modality recognition. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 181–189. Springer (2019)

[41] Robinson, C., Trivedi, A., Blazes, M., Ortiz, A., Desbiens, J., Gupta, S., Dodhia, R., Bhatraju, P.K., Liles, W.C., Lee, A., et al.: Deep learning models for covid-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement. medRxiv (2021)

[42] Robinson, D.A.: The oculomotor control system: A review. Proceedings of the IEEE **56**(6), 1032–1049 (1968)

[43] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. arXiv: Computer Vision and Pattern Recognition (2014)

[44] Saab, K., Hooper, S.M., Sohoni, N.S., Parmar, J., Pogatchnik, B.P., Wu, S., Dunnmon, J., Zhang, H., Rubin, D.L., Ré, C.: Observational supervision for medical image classification using gaze data. In: Medical Image Computing and Computer-Assisted Intervention (2021)

[45] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)

[46] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. arXiv preprint arXiv:2201.09873 (2022)

[47] Shen, X., Lam, W.: Towards domain-generalizable paraphrase identification by avoiding the shortcut learning. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 1318–1325 (2021)

[48] Shome, D., Kar, T., Mohanty, S.N., Tiwari, P., Muhammad, K., AlTameem, A., Zhang, Y., Saudagar, A.K.J.: Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. International Journal of Environmental Research and Public Health **18**(21), 11086 (2021)

[49] Shorfuzzaman, M., Hossain, M.S.: Metacovid: A siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients. Pattern recognition **113**, 107700 (2021)

[50] SIIM-ACR pneumothorax segmentation (2020), [online] Available: https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

[51] Swensson, R.G.: A two-stage detection model applied to skilled visual search by radiologists. Attention Perception & Psychophysics **27**, 11–16 (1980)

[52] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)

[53] Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging (2022)

[54] Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) **53**(3), 1–34 (2020)

[55] Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. arXiv preprint arXiv:2006.09994 (2020)

[56] Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 579–588 (2021)

[57] Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine **15**(11), e1002683 (2018)

[58] Zhang, Z., Zhang, H., Zhao, L., Chen, T., Pfister, T.: Aggregating nested transformers. arXiv preprint arXiv:2105.12723 (2021)

[59] Zuiderveld, K.: Contrast limited adaptive histograph equalization. Graphic Gems IV p. 474–485 (1994)