Rectify ViT Shortcut Learning by Visual Saliency

Chong Ma, Lin Zhao, Yuzhong Chen, Lei Guo, Tuo Zhang, Xiuntao Hu, Dinggang Shen, Fellow, IEEE, Xi Jiang, and Tianming Liu, Senior Member, IEEE

Abstract—Shortcut learning in deep learning models occurs when unintended features are prioritized, resulting in degenerated feature representations and reduced generalizability and interpretability. However, shortcut learning in the widely used vision transformer (ViT) framework is largely unknown. Meanwhile, introducing domain-specific knowledge is a major approach to rectifying the shortcuts that are predominated by background-related factors. For example, eye-gaze data from radiologists are effective human visual prior knowledge that has the great potential to guide the deep learning models to focus on meaningful foreground regions. However, obtaining eye-gaze data can still sometimes be time-consuming, labor-intensive, and even impractical. In this work, we propose a novel and effective saliency-guided ViT (SGT) model to rectify shortcut learning in ViT with the absence of eye-gaze data. Specifically, a computational visual saliency model (either pretrained or fine-tuned) is adopted to predict saliency maps for input image samples. Then, the saliency maps are used to filter the most informative image patches. Considering that this filter operation may lead to global information loss, we further introduce a residual connection that calculates the self-attention across all the image patches. The experiment results on natural and medical image datasets show that our SGT framework can effectively learn and leverage human prior knowledge without eye-gaze data and achieves much better performance than baselines. Meanwhile, it successfully rectifies the harmful shortcut learning and significantly improves the interpretability of the ViT model, demonstrating the promise of transferring human prior knowledge derived visual saliency in rectifying shortcut learning.

Index Terms—Interpretability, saliency, shortcut learning, vision transformer (ViT).

I. INTRODUCTION

As deep learning has been increasingly adopted in natural language processing (NLP), computer vision (CV), and medical image analysis [39], among others, the black box issue that comes with it has attracted extensive attention and discussion. Especially when deep learning models step into risk-sensitive scenarios, such as autonomous driving or medical diagnostics in real-world applications, questions about their interpretability and transferability/generalizability have become more intensive. Many researchers have come to believe that shortcut learning, as an implicitly harmful phenomenon, is a major reason causing the poor robustness and low transferability/generalizability of deep learning models [22] when applying to new scenarios or datasets.

Shortcut learning [17], [22] refers to the phenomenon that models highly rely on unintended or nonrobust features as shortcuts, failing to learn robust features and capture high-level semantic understanding and reasoning. Recent work [47] revealed that background-related factors in natural images can be representative and predominant shortcuts that drastically affect the model’s performance. Similar problems also exist in the medical image analysis field and may cause even more harmful [46], [56], [79] outcomes when training deep models for clinical diagnosis. In addition, among different strategies aiming to address shortcut learning, domain-specific knowledge is considered an important approach to avoiding the unintended representations/features learned by shortcut learning [17], [22], [57]. For example, human eye-gaze reflects the area in the image that they are viewing, so foreground and background information can be distinguished based on the eye-gaze heatmap generated from eye-gaze data. In natural images [58], eye-gaze data can be used to guide fine-grained image classification. In medical images [73], radiologists’ eye-gaze data can be particularly helpful in assisting disease diagnosis. Eye-gaze from radiologists when examining medical images, as a direct expression of human subjective knowledge or preference, can capture the subtle gaze behavior and, therefore, identify the areas that are potentially

Manuscript received 19 December 2022; revised 18 July 2023; accepted 24 August 2023. Date of publication 13 September 2023; date of current version 3 December 2024. The work of Xi Jiang was supported in part by the NSFC under Grant 61976045 and Grant 62276050, in part by the Sichuan Science and Technology Program under Grant 2021YJ0247, and in part by the Special Fund for Basic Scientific Research of Central Colleges under Grant ZYGX2021J036. The work of Dinggang Shen was supported by the NSFC under Grant 62131015. The work of Xintao Hu was supported by the NSFC under Grant 62076205. The work of Tuo Zhang was supported by the NSFC under Grant 31971288, Grant U1810265, and Grant 31671005. The work of Lei Guo was supported by the National Natural Science Foundation of China (NSFC) under Grant 61936007. The work of Chong Ma and Lin Zhao contributed equally to this work. (Corresponding authors: Xi Jiang; Tianming Liu.)
more informative and related to the medical diagnosis tasks. Although it is much easier to obtain eye-gaze data than to obtain fine-grained manual annotations, this still requires time and labor, and in some cases, we cannot even obtain any real auxiliary data.

In computational vision, saliency information characterizes the importance of different features in an image [2], and it can effectively mark the regions of the image that contains higher semantic information. In addition, the training of the saliency prediction model often relies on human annotation, such as mask [78], bounding box [43], and eye-gaze [3]. Therefore, using the pretrained or fine-tuned saliency model not only provides prior knowledge learned from humans but also serves as domain knowledge to assist the model in accomplishing some specialized tasks, e.g., the saliency model trained with the radiologists’ eye-gaze data can imitate the radiologists to diagnosis important regions (possible lesion regions) in the medical images, which, to some extent, alleviates the reliance on refined human annotation. More importantly, we believe that using prior knowledge provided by the saliency model to assist the model training can alleviate the shortcut learning problem, thus further improving the performance of the model.

In the field of CV, vision transformer (ViT) [15] has been receiving increasing attention and gradually become the fundamental model [8], [12], [72], [77], [82]. Many works have improved and extended ViT, such as DeiT [70], Swin transformer [44], MAE [23], and MoCo-v3 [7]. Most of the recent large-scale models, such as SAM [36] and CLIP [55], are also employed ViT as a feature extraction module. However, few works investigated the shortcut learning problem in ViT models. Even though it is possible to pretrain a generalized ViT model using large amounts of data, the lack of data is still an obvious problem when it comes to specific downstream tasks. In response, in this work, we reveal that ViT models have serious shortcut learning problems, and we propose a novel and effective saliency-guided ViT (SGT) model to alleviate the influence of shortcut learning when the domain knowledge, such as eye-gaze data, is not available, as shown in Fig. 1. Specifically, our SGT model first learns a saliency map from natural or medical image datasets with eye-gaze data, that is, we use real eye-gaze data to fine-tune a saliency prediction model to predict the saliency map of images based on human prior knowledge. Then, we introduce a mask encoder that uses the mask generated from the saliency map to distill the most informative regions from the data without eye-gaze information and perform self-attention learning only on these regions. Our saliency prediction model is trained on both natural and medical image datasets with eye-gaze data for the sake of generating valid saliency maps even on datasets without eye-gaze information. Moreover, in order to maintain the global information of the image, we add the global residual connection (RC) at the last transformer encoder layer. In this way, our SGT model can learn more effective features from prior visual information and significantly reduce the shortcut learning phenomenon due to dataset size or bias problems, as shown in Fig. 1. We evaluate our SGT model on four independent public datasets (including two natural and two medical image datasets): FIGRIM [5] and CAT2000 [3] in natural images and INbreast [52] and SIIM-ACR (the dataset of 2019 Pneumothorax Segmentation Challenge [65]) in medical images. Our experiments show that the proposed SGT model significantly improves image classification and medical diagnosis accuracy, and successfully avoids the harmful shortcut learning phenomenon compared with the baseline ViT models and other popular methods.

Overall, the main contributions of this work are given as follows.

1) We reveal that the shortcut learning phenomenon is prevalent in the ViT framework and quantitatively evaluate the severity of shortcuts in some representative ViT baseline models.

2) We propose a novel SGT model that uses a visual saliency mask operation to guide the model to focus on the regions with relevant and discriminative features.

3) Our proposed mask encoder and global RC effectively learn the informative information in the image without losing the global information as well. We can apply this strategy to various ViT models, which improves the robustness and interpretability of the models.

II. RELATED WORK

A. Shortcut Learning

Recently, shortcut learning has received much attention in deep learning areas, such as CV [4], [13], [47], [48], [56], [83]
and NLP [11], [17], [18], [51], [53], [64], [69]. For most of the tasks in deep learning, both the training and test sets come from the same dataset. This results in some implicit biased information (like cows that often appear with the grass and birds that often appear with the sky) in the database that exists in both the test and training sets. Thus, the model may learn the background feature instead of the real object, and this kind of shortcut is also hard to find through testing metrics [22]. More recently, Tang et al. [69] uncovered the propensity of large language models (LLMs) to leverage shortcuts within prompts for downstream tasks, and the high-quality prompts do not alleviate the impact of shortcut learning. This further demonstrates the importance of studying shortcut learning problems.

Many methods have been devised to mitigate the negative effects of shortcuts. For example, Luo et al. [47] proposed a framework named COSOC to extract the foreground objects in images without extra supervision to avoid harmful background shortcuts in few-shot learning (FSL). Similarly, Chen et al. [9] proposed a guidance strategy to focus on foreground objects in images in FSL. Robinson et al. [57] found that the shortcut present in the contrastive learning task suppresses important feature learning. They modified the positive and negative samples by implicit feature modification (IFM) to mitigate this phenomenon. Zhong et al. [83] proposed a data augmentation strategy to suppress the existing shortcut features, and they also used adaptive pruning and knowledge distillation methods to further mitigate the shortcut learning problem of the model. In the medical imaging field, Robinson et al. [56] used feature disentanglement in a multitask training framework to prevent shortcut learning in the context of automated classification of COVID-19 chest X-ray images. Mahapatra et al. [48] have proposed an interpretability-guided inductive bias approach that enforces learned features to yield more distinct and spatially consistent saliency maps for different class labels of trained models, which leads to improved model performance and reduces the shortcut learning problem. In contrast to the above approach, we believe that introducing human prior knowledge and using it to guide the deep learning network to focus on more important and relevant regions can be effective in avoiding the shortcut learning phenomenon.

B. Human Eye-Gaze

Eye-gaze data can be directly obtained by using eye tracking methods [19] and showing the interest area by humans in image or other stimuli. In CV, many works [32], [38], [58], [62] used eye-gaze data directly as auxiliary information to perform tasks such as image classification, video action recognition, or image segmentation, and thus, they demonstrated the potential of eye-gaze data to replace the expensive manual annotation in deep learning. For example, Lai et al. [38] used eye-gaze data to accomplish salient target segmentation, video action classification, and fine-grained image classification, respectively. Liu et al. [40] suggest that the eye-gaze information of experts in a specific field can not only guide novices but also be used to guide neural network models. The method incorporates eye-gaze into a GAN network and significantly improves model performance. They also observed that the model guided by eye-gaze data pays more attention to task-relevant regions, thereby alleviating the problem of shortcut learning. In the literature, many works [33], [34], [49], [50], [67], [73] have demonstrated the important role of eye-gaze data in medical image analysis as well. For instance, Wang et al. [73] used the radiologists’ eye-gaze data to generate the attention map and used it to constrain the model’s attention. Stember et al. [67] developed a CNN-based segmentation method using the information from the radiologists’ gaze on the lesion. Khosravan et al. [33], [34] segmented the disease area using the radiologists’ eye-gaze information. In the domain of cognitive neuroscience, Gao et al. [21] successfully employed the integration of eye-gaze patterns with brain functional connectivity to forecast cognition-related phenotypic scores. Although capturing eye-gaze data is less time-consuming than fine-grained manual labeling, there are still some disadvantages that make the capture step complicated and unfriendly, especially for radiologists in medical image diagnosis, and thus, the capture process becomes slow and time-consuming. Therefore, we designed an eye-gaze collection system specifically for radiologists, which greatly saves the collection time and lays the foundation for our subsequent work. This system can also be extended to other situations for eye-gaze collection tasks. The source code of the collection system is publicly available at Github, and more details can be found in the Supplementary Material.

C. Visual Saliency Prediction

A practical and alternative way is to use the saliency model to predict saliency maps that attract human visual attention. Nowadays, there are two main tasks in artificial saliency: saliency prediction and salient object detection (SOD). The saliency prediction model outputs a Gaussian distributed heatmap for direct simulation of human fixation information. For example, many works [10], [16], [26], [27], [28], [37], [80] use CNN-based models to generate the saliency map of images. Among them, EML-NET [27] performs well on several benchmarks [1]. UnisalNet [16] combines MobileNet [61] with a decoder based on RNN [71] to output predicted saliency maps. Recent method TranSalNet [45] combines a transformer encoder with DenseNet [25] and, finally, uses a CNN decoder to output predicted saliency maps. In comparison, the SOD model outputs a mask map, which simulates human visual perception in locating the most significant object(s) in a scene. Many works such as [41], [74], [76], [81], and [20] used fully convolutional network (FCN)- or VGG-based model to generate the salient object map. Recently, the VST [42] model achieved excellent prediction results on both RGB and RGB-D datasets based on the transformer architecture. Zhuge et al. [84] proposed a novel integrity cognition network ICON to detect salient objects by leveraging integral features at both microlevels and macrolevels and also achieved excellent results. In order to be independent of the eye-gaze data and to obtain the guidance mask, we use a saliency prediction

1https://github.com/MoMarky/eye-tracking-system-for-radiologists
We implemented this idea by introducing a saliency mask on the model to focus on the most important and relevant regions. For this, we use visual saliency to guide the model to maintain the correlations of all patches. Specifically, the input image can be divided into N patches, and the ViT model maps the image patches $x_i^l$ ($i = 1, 2, \ldots, N$) to D dimension embedding with a trainable linear projection $E \in \mathbb{R}^{H \times W \times C \times D}$, where $H$, $W$, and $C$ are the height, weight, and channel of the images. After patch embedding, we obtain the input $z$ for the SGT model

$$z_{in} = [x_{class}; (x_p^1 E; x_p^2 E; \ldots; x_p^N E)] + E_{pos}$$

where $x_{class} \in \mathbb{R}^N$ is the class token for classification, $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ is the learnable position embedding, and $(x_p^1 E; x_p^2 E; \ldots; x_p^N E)$ are the original patch embeddings. After $K$ transformer encoder layers, we obtain the layer output $z_K$

$$z_{K} = [x_{class}; (e_K^1; e_K^2; \ldots; e_K^N)]$$

where $e_K^1; e_K^2; \ldots; e_K^N$ are the encoded patch embeddings after $K$ encoder layers and $K$ is the number of transformer encoder layer. Then, we directly use the mask generated by the predicted saliency map to filter out the most important and relevant patches, and obtain the output $z_L$ by the mask encoder

$$z_{MASK} = [x_{class}; (e_K^1; e_K^2; \ldots; e_K^N) \odot \text{Mask}]$$

$$z_L = \begin{cases} 
\text{MSA}(\text{LN}(z_{MASK})) + z_{MASK}, & \text{if } l = 1 \\
\text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, & \text{if } l = 2, \ldots, L 
\end{cases}$$

where Mask $\in \mathbb{R}^N$ in (3) is the binary saliency mask detailed in Section III-C; MSA and LN are the multiheaded self-attention operation. In order to focus on important image regions without losing global information, we introduce a mask encoder to the vanilla ViT architecture, which directly uses the mask generated by the image saliency to filter out important information and process only this information in the subsequent encoder layer. Before the final transformer encoder layer of the SGT, we added the remaining embedding patches after mask operation back to retain global information and maintain the correlations of all patches.

In this work, to avoid the ViT model from learning irrelevant or harmful shortcuts, we use visual saliency to guide the model to focus on the most important and relevant regions. We implemented this idea by introducing a saliency mask on input image patches and an additional global RC on the SGT model.

The architecture of our proposed SGT model is shown in Fig. 2. First, we use the eye-gaze data (yellow points) collected on the image as ground truth to train a saliency model and processed to generate a saliency mask. Meanwhile, the global embeddings of the image are obtained after the input image goes through the patch embedding and $K$ transformer encoder layers. After that, we introduce a mask encoder with $L$ encoder layers, which directly uses the saliency of the image to filter out the important patch embeddings (highlighted by the red rounded rectangle) and only perform self-attention learning only on them. After the mask encoder, we applied global RC before the last transformer encoder layer (highlighted by green arrow). Finally, we use the class token as the representation feature to perform the classification task.
self-attention and layer norm; and $L$ is the number of mask encoder layers. After the mask operation described in (3), we obtain the filtered important local embeddings $z_{\text{MASK}}$. Then, $z_{\text{MASK}}$ is input to the mask encoder, as shown in Fig. 2. Specifically, as shown in (4), all $L$ layers of the mask encoder process only the previously filtered embeddings. Then, as shown in (5), we add back the rest of the embeddings dropped earlier before the last transformer encoder layer of the model and use the class token in the final layer’s output as the input of the classifier to obtain the output of the model

$$z_{\text{res}} = [x_{\text{class}}; z_L \cup \tilde{z}_K]$$
$$z_{\text{out}} = \text{MSA}(\text{LN}(z_{\text{res}})) + z_{\text{res}} = [x_{\text{class}}; e_{\text{res}}^{1:N}]$$
$$\text{output} = \text{MLP}(\text{LN}(x_{\text{class}}))$$

where $z_L$ is the embeddings after mask encoder, $\tilde{z}_K$ is the embeddings of the remaining positions of $z_K$ after the mask operation shown in (3), $e_{\text{res}}^{1:N}$ is the patch embeddings of final transformer encoder layer, and MLP is the multilayer perceptron. In this method, $K$ and $L$ are preset parameters, and detailed ablation studies are presented in Section IV-D.

B. Human Attention Learning

In order to better learn human attention, we choose a saliency prediction model. Specifically, we use a computational visual saliency model EML-NET [27] to predict the saliency heatmap. EML-NET is a scalable encoder–decoder framework that leverages multiple deep CNN models (such as VGG-16 [66], ResNet-50 [24], DenseNet-161 [25], and NasNetLarge [85]) to better extract visual features for saliency prediction. It uses a combined loss function consisting of three saliency metrics, including the Kullback–Leibler divergence (KLD) [6], the modified Pearson’s correlation coefficient [27], and normalized scanpath saliency [27]. In our study, due to the fact that EML-NET was trained using the SALICON dataset [29], the largest dataset for saliency prediction in natural images, we directly use the pretrained EML-NET to generate saliency maps for the natural image datasets (FIGRIM [5] and CAT2000 [3]). For the medical image datasets (INbreast [52] and SIM-ACR [65]), the pretrained EML-NET is fine-tuned using the accompanied eye-tracking recording data. In addition, we conducted a comparative analysis of two saliency prediction models (UnisalNet [16] and TranSalNet [45]) and one saliency detection model (VST [42]) mentioned in previous Section II-C. Following the aforementioned procedure, we obtained fine-tuned models and employed the generated saliency maps to guide our SGT model.

In this way, we can generate effective saliency maps for both natural and medical images regardless of whether the image has corresponding eye-gaze data or not. As a comparison study, we also compared the proposed SGT model with the saliency maps generated in four different ways, that is, the one predicted by the EML-NET, the one predicted by an SOD model [42], the one generated directly using the eye-gaze recordings, and finally, the one generated by model itself using Grad-CAM [63] method. All comparison results are shown in Section IV-D2.

C. Mask Encoder

In this work, we propose a mask encoder that directly uses the saliency mask to filter out the most important and relevant regions of the image for the model training. As shown in Fig. 3, we first use the saliency model introduced in Section III-B to predict the saliency map of the images. Then, we sort the values in the saliency map and select the top $M$ values as the important regions, and the binary saliency mask is obtained by setting 1 to these regions and 0 to the rest. Here, we select the top 20%, 25%, 30%, and 40% regions in the saliency map, meaning that we mask out 80%, 75%, 70%, and 60% of the unimportant areas of the image. The detailed ablation study can be found in Section IV-D2. As the patch embedding layer in ViT model maps the input image to $14 \times 14 \times D$ (suppose that the image size is $224 \times 224$ and the patch size is $16 \times 16$), we also resize the saliency mask to $14 \times 14$. Meanwhile, as shown in Fig. 3, we reshape the input embeddings except for the class token (brown squares) into $14 \times 14 \times D$ 3-D features. Then, the 3-D features are multiplied directly with a resized saliency mask to obtain the distilled embeddings. We then flatten these distilled embeddings, add the previous class token, and use these embeddings as input for the subsequent transformer encoder layers so that the mask encoder only learns from these features of informative regions. We believe that, in this way, the sensitivity of the model to important regions can be improved.

D. Random Guidance Strategy

In fact, human gaze information is not always effective for the classification task, especially for aimless observation tasks in natural images. As shown in Fig. 4, the first column represents human attention, which shows that humans tend to unconsciously focus their attention on the characters when freely watching these scenes. However, in some scenes, the semantic information of the characters cannot be used to distinguish the scene, so human attention in these cases is biased prior information. The second column in Fig. 4 shows the prediction results of the saliency model trained on these eye-gaze data, which are almost consistent with human attention. This demonstrates the saliency model’s advantage in learning prior
knowledge but also exposes its weakness in learning human biases. In our SGT model, using biased prior knowledge (such as the heatmaps in the second column of Fig. 4) leads the model to focus on the wrong facial regions instead of scene-related regions, such as beds and windows. This causes the model to associate incorrect semantic information with scene information, resulting in a decrease in model performance.

To address the issue of bias in prior information itself, we propose a simple yet highly effective strategy: random guidance. Specifically, during the training of our SGT model, we hypothesize that balancing the model’s self-learning with the guidance of prior knowledge can further improve model performance while avoiding potential biases caused by human learning errors. We believe that transformer models have the ability to locate important regions even without relying on prior information (as demonstrated by the ViT baseline in Section IV-D2) although their performance still needs to be improved. To implement the random guidance strategy, we define a threshold and randomly use the guidance of prior information during the training process. A detailed description is provided in Section IV-D3. As shown in the third column of Fig. 4 and experimental results in Section IV-D3, after using the random guidance strategy, the SGT model successfully focuses on the regions truly related to the scene rather than the regions previously focused on by the human attention, demonstrating the effectiveness of our random guidance strategy.

IV. EXPERIMENTS

A. Datasets

We evaluate our SGT model on four different datasets, including two natural image datasets, FIGRIM [5] and CAT2000 [3], and two publicly available medical image datasets, INbreast [52] and SIIM-ACR [65]. Among them, FIGRIM [5] provides eye fixation data for a total of 2787 images spanning 21 indoor and outdoor scene categories. CAT2000 [3] has 2000 training images and 2000 testing images from 20 different categories with eye-tracking data. The SIIM-ACR [65] is a chest X-ray dataset with only pneumothorax disease, and recently, the work [60] randomly selected 1170 images, with 268 cases of pneumothorax, and collected gaze data from three radiologists. The INbreast [52] dataset includes 410 full-field digital mammography images that were collected during low-dose X-ray irradiation of the breast. However, there is no corresponding eye-gaze data in INbreast [52] dataset. Thus, we invited a radiologist with ten years of experience in diagnosing breast cancers and collected the eye-gaze data. Detailed analysis and presentation of eye-gaze collection and more data preprocessing can be found in the Supplementary Material.

B. Evaluation Metrics

We adopt F1-score (F1) and area under curve (AUC) evaluation metrics to evaluate our model performance comprehensively and Grad-CAM [63] to visualize the attention heatmap of models on the image. To test the interpretability of the model, we first evaluated the similarity between the model’s own attention and the actual saliency of the images during the testing phase using the structure similarity index measure (SSIM) [75] score. During the testing phase, we employed a vanilla ViT model to load the previously trained parameters for predicting the test images. Furthermore, we also employed the Grad-CAM [63] method to obtain the attention heatmap of the model while obtaining the classification results. We posit that the model’s attention guided by prior knowledge indicates a stronger focus on the truly salient regions. During SSIM score computation, we first resized the saliency map of the saliency model and the attention map of the SGT model to the same size. Subsequently, we normalized them and fed them into the SSIM function to obtain the similarity score between the two heatmaps. This metric can evaluate the similarity between the two images from three aspects: structure, brightness, and contrast. Its score ranges between 0 and 1, with a higher score indicating greater similarity.

Furthermore, we present a new evaluation metric for assessing shortcut learning problems. In Section I, we demonstrate that shortcut learning in models is mainly caused by their incorrect attention to irrelevant regions in the image. Conversely, when a model exhibits less severe shortcut learning, it concentrates more on informative regions in the image. Based on this observation, we propose a new evaluation metric. Specifically, for a trained model, we conduct round-one testing and record classification performance metrics (such as Acc, AUC, or F1 score) and the model’s attention heatmaps for each image. Subsequently, we employ Gaussian noise to cover the areas of the image that the model primarily focuses on and perform round-two testing while recording the same performance metrics. As previously mentioned, a model with less severe shortcut learning can focus better on critical regions in the image. Thus, if we cover the areas that the model focuses on during the round-two testing, we should observe a significant decline in performance metrics. Therefore, we calculate a metric for evaluating shortcut learning problems by subtracting the metric values obtained in the round-two testing...
from those obtained in the round-one. We define this as the difference after covering (DAC) metric. The calculation process is shown in Fig. 5, where the top half represents the round-one testing, and the bottom half represents the round-two testing. First, we calculate the bounding box based on the attention map of the model from the round-one testing. Next, we fill this area with Gaussian noise and conduct the round-two testing, with the same model used in both rounds. Finally, the value of the DAC metric is obtained as Result_1 – Result_2. In this experiment, we choose the F1 score as the value for Result_1 and Result_2. The larger DAC values represent models with less severe shortcut learning problems.

C. Implementation Details

For all models in our experiment, we use the weights pretrained on ImageNet [14] and fine-tune each dataset. Specifically, we fine-tune all models, including ResNet [24], EfficientNet [68], Swin transformer [44], ViT baselines, and our SGT model, using Adam [35] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a batch size of 64. We train all models with 60 epochs with an initial learning rate of $1 \times 10^{-4}$, a weight decay of $1 \times 10^{-6}$, and 5 epochs of warm-up, and we use the cross-entropy loss with label smoothing hyperparameter of 0.1 as the final loss function. All saliency maps were extracted by our saliency prediction model, and all images were simply resized to $224 \times 224$ pixels. Note that the saliency map is the attention map of the model from the round-one testing. Next, we fill this area with Gaussian noise and conduct the round-two testing, with the same model used in both rounds. Finally, the value of the DAC metric is obtained as Result_1 – Result_2. In this experiment, we choose the F1 score as the value for Result_1 and Result_2. The larger DAC values represent models with less severe shortcut learning problems.

D. Main Property Studies

We ablative our SGT in three aspects of model architecture, mask type, and random guidance strategy. Several interesting properties are observed.

1) Model Architecture: In the ViT framework, as the depth of encoder layer increases, its attention distance [15] also changes, so it is necessary to discuss the depth ($L$ in Fig. 2) and position ($K$ in Fig. 2) of the mask encoder on the final results. We also consider the comparison results of the global RC in the last transformer encoder layer. In Table I, the first three columns are the values of $K$, $L$, and the flag of global RC, respectively. In the RC column, $\checkmark$ means the model with the global RC and $\times$ means without. We use ViT-S (with 12 transformer encoder layers) as the baseline and compare ten model architectures on the INbreast [52] and FIGRIM [5] datasets, and we adopt AUC, SSIM score, and DAC as metrics.

As shown in Table I, the best results appear in using a seven-layer mask encoder placed at the fifth layer of the model. Also, using global RC has a significant improvement in the results. Experiment results in Table I show that the ViT model may learn shortcut knowledge more easily at intermediate layers, and applying guidance directly to layers in intermediate locations would greatly alleviate the problem. Therefore, the input image is first encoded by four transformer encoder layers and then learned by a seven-layer mask encoder for important regions, and finally, the global information is complemented by RC. Since we are using ViT-S [15] as the baseline, we call the above model architecture SGT-S.

2) Different Types of Masks: We considered the effect of different types of masks and conducted comparative experiments on INbreast [52] and FIGRIM [5] datasets using the SGT-S model. Each row of the mask type in Table II represents (from top to bottom) the baseline without the mask, the human attention mask, the mask generated by Grad-CAM [63] of the model, the mask generated by an SOD model VST [42], and the mask generated by two saliency prediction models UnisalNet [16] and TranSalNet [45], and the last eight rows are four types of degrees of random mask and our proposed saliency mask generated by EML-Net [27], respectively.

As shown in Table II, in the INbreast [52] dataset, the saliency mask predicted by TranSalNet performed the highest AUC metric, followed by the real eye-gaze mask and the EML-Net mask (SGT-S 75% Mask). This indicates that the use of more powerful saliency models can indeed enhance the overall performance of the model. In addition, it confirms that our random guidance strategy further optimizes the guidance of prior information and model self-learning, surpassing the results of the eye-gaze mask. However, in the

![Fig. 5. Calculation process of DAC metric.](image-url)
and reaches its peak at $T = 0.6$. In natural images, the result at $T = 0.6$ is still higher than others. It is worth noting that the result at $T = 0.0$ is higher than those at $T = 0.2, 0.4, 0.5, 0.8$, further proving our previous idea that, when using the ViT model pretrained on ImageNet as the baseline, the model already has the ability to identify important regions. Excessive use of biased prior information can adversely affect model performance. It can also be observed in Fig. 4 that some of the bias issues present in the prior information can also be balanced by learning the model itself. In summary, our random guidance strategy can effectively balance the model’s self-learning with the guidance of prior information. In the above ablation experiments, if not specified, the default parameters of the model are $K = 4$, $L = 7$, $T = 0.6$, and mask degree = 75%.

### E. Comparisons With Baseline

In this article, we compare the experiment results of SGT and ViT model on four datasets, including two medical image datasets (INbreast [52] and SSIM-ACR [65]) and two natural image datasets (FIGRIM [5] and CAT2000 [3]). We also compare two scales of ViT and SGT model, small and base, where patch size, layer number, and embedding length are 16, 12, and 384 in the small scale and 16, 12, and 768 in the base scale. All ViT and SGT models were pretrained on the ImageNet dataset [14] and fine-tuned on each dataset in our experiment. Fig. 6 shows some comparison examples on INbreast [52] and FIGRIM [5] datasets, and Table IV shows the AUC, F1, SSIM, and DACC metrics for these models on two natural image datasets and two medical image datasets. In Fig. 6(a), the ViT-S model without the guidance of saliency tends to focus on the background region in mammogram images, as the big difference between breast tissue and background. We also found that the ViT-S model often incorrectly focuses on the nipple region, as shown in the fifth column of Fig. 6(a). This is because there are no nipple-related diseases in the INbreast [52] dataset and the model focuses on it only because of its characteristic appearance. In contrast, our proposed SGT-S model corrected this shortcut learning phenomenon, as shown in the bottom two rows in Fig. 6(a). We also found that, in the first two columns of Fig. 6(b), since the “skyscraper” often appears together with the sky, the ViT-S model tends to focus on the more distinguishable sky areas. With our proposed saliency guidance, the SGT-S model also

| Mask Type | INbreast [52] | FIGRIM [5] |
|-----------|---------------|------------|
| VIT (Baseline) | 18.79 | 0.47*** | 1.81 | 99.47 | 0.33*** | 0.24 |
| Eye-gaze | 18.20 | 0.33*** | 1.64 | 99.07 | 0.21*** | 0.31 |
| Grad-CAM [63] | 20.48 | 0.71*** | 1.64 | 99.64 | 0.38*** | 0.26 |
| VST [42] | 22.53 | 0.48*** | 2.89 | 99.45 | 0.36*** | 0.30 |
| UltraSalNet [14] | 20.21 | 0.51*** | 3.13 | 99.55 | 0.51*** | 0.27 |
| TransSalNet [65] | 24.60 | 0.70*** | 3.75 | 99.68 | 0.52*** | 0.33 |
| Random (80% Mask) | 20.17 | 0.36*** | 4.82 | 99.29 | 0.38*** | 0.24 |
| Random (75% Mask) | 91.08 | 0.38** | 5.70 | 99.33 | 0.31*** | 0.24 |
| Random (60% Mask) | 90.65 | 0.43*** | 4.93 | 99.32 | 0.30*** | 0.25 |
| SSTG (80% Mask) | 90.63 | 0.39** | 4.29 | 99.18 | 0.24** | 0.21 |
| SSTG (75% Mask) | 90.63 | 0.69** | 7.57 | 99.69 | 0.58** | 0.31 |
| SSTG (60% Mask) | 91.38 | 0.41** | 6.47 | 99.66 | 0.57** | 0.29 |

** *: p < 0.001, **: p < 0.01, ***: p < 0.025, ns: p > 0.025**

| Threshold | INbreast [52] | FIGRIM [5] |
|-----------|---------------|------------|
| T = 0.0  | 18.79 | 0.47*** | 1.81 | 99.47 | 0.33*** | 0.24 |
| T = 0.2  | 19.90 | 0.47*** | 5.79 | 99.24 | 0.50** | 0.27 |
| T = 0.4  | 20.18 | 0.52*** | 6.45 | 99.15 | 0.50*** | 0.31 |
| T = 0.5  | 20.62 | 0.52*** | 7.12 | 99.26 | 0.53*** | 0.30 |
| T = 0.6  | 23.63 | 0.74 | 7.80 | 99.84 | 0.61 | 0.34 |
| T = 0.8  | 22.53 | 0.54** | 7.80 | 99.26 | 0.48** | 0.32 |

** **: p < 0.001, ***: p < 0.01, *: p < 0.025, ns: p > 0.025**

The EML-Net mask performed optimal results, surpassing the results of both the eye-gaze and TranSalNet mask. This further indicates our conclusion in Section III-D that eye-gaze data in natural images contain human bias. The utilization of stronger saliency models also learns stronger human bias, resulting in inferior performance compared to EML-Net. For the Grad-CAM [63] mask, the guidance of the model comes entirely from the model itself, and although it will focus on some regions in the ongoing iterations, it is likely to be shortcut regions rather than really useful regions. For the VST [42] mask, the guidance of the model comes from an SOD model, which outputs a segmentation mask of the salient object. The mask used for segmentation is often along the boundary and contains all the range of the object. This is not appropriate for us because we want to get the most salient small area rather than the whole of an object, and a detailed analysis can be found in the Supplementary Material. For the random mask, we tried four scales and found that most of the metrics were better than the baseline. We indicate that there are more redundant features in images, and the baseline model will be affected by these redundant features, thus creating a shortcut learning problem that leads to performance degradation. In all comparison results, we adopted EML-Net (SGT-S 75% Mask) as the foundational model for human attention learning, considering its superior performance on the FIGRIM and INbreast datasets, as well as its simple model architecture.

3) Random Guidance Strategy: To balance the prior knowledge with the model’s own learning, we introduce a random guidance strategy, and this strategy could also mitigate the possible negative effects of human prior knowledge. Specifically, we first set a threshold value $T \in [0, 1]$. Then, our SGT model generates a random probability $p \in [0, 1]$ in each batch of training, and the model uses the saliency mask guidance if $p \geq T$ and not if $p < T$. As shown in Table III, we select five thresholds ($T = 0.2, 0.4, 0.5, 0.6, 0.8$) and one baseline test ($T = 0.0$) for comparison experiments on the INbreast [52] and FIGRIM [5] datasets, respectively, based on our SGT-S model. It can be seen that, without using the random strategy, SGT performs poorly on the medical image dataset, but, after using the random strategy, the model performance steadily improves and reaches its peak at $T = 0.6$. In natural images, the result at $T = 0.6$ is still higher than others. It is worth noting that the result at $T = 0.0$ is higher than those at $T = 0.2, 0.4, 0.5, 0.8$, further proving our previous idea that, when using the ViT model pretrained on ImageNet as the baseline, the model already has the ability to identify important regions. Excessive use of biased prior information can adversely affect model performance. It can also be observed in Fig. 4 that some of the bias issues present in the prior information can also be balanced by learning the model itself. In summary, our random guidance strategy can effectively balance the model’s self-learning with the guidance of prior information. In the above ablation experiments, if not specified, the default parameters of the model are $K = 4$, $L = 7$, $T = 0.6$, and mask degree = 75%.
Fig. 6. Comparison examples on (a) INbreast and (b) FIGRIM datasets. We choose the ViT-S and SGT-S models to demonstrate the improvement of our proposed method for the shortcut learning phenomenon. Each column corresponds to the same example. In each column, we use a pair of figures to show the attention of the model, where the top half is the attention heatmap (top) and the activation region (bottom) for the ViT-S model, and the bottom half is the attention heatmap (top) and the activation region (bottom) for the SGT-S model.

Table IV

| Dataset | Metrics | Small(16,12,384) | Base(16,12,768) |
|---------|---------|-----------------|-----------------|
|         |         | ViT-S [15]     | SGT-S           | ViT-B [15]     | SGT-B           |
| INBreast [52] | AUC↑     | 88.79           | 93.63           | 90.18           | 96.91           |
|         | F1↑      | 91.01           | 91.77           | 90.45           | 92.71           |
|         | SSIM↑    | 0.477           | 0.74            | 0.496           | 0.77            |
|         | DAC↑     | 1.81            | 7.80            | 2.16            | 8.94            |
| SIM-ACR [53] | AUC↑     | 77.53           | 85.53           | 81.50           | 85.46           |
|         | F1↑      | 82.67           | 85.36           | 84.92           | 85.32           |
|         | SSIM↑    | 0.296           | 0.65            | 0.276           | 0.44            |
|         | DAC↑     | 3.55            | 6.13            | 4.13            | 6.97            |
| FIGRIM [5] | AUC↑     | 99.47           | 99.84           | 99.77           | 99.81           |
|         | F1↑      | 95.13           | 94.84           | 94.94           | 95.24           |
|         | SSIM↑    | 0.63            | 0.64            | 0.60            | 0.58            |
|         | DAC↑     | 0.24            | 0.34            | 0.26            | 0.36            |
| CAT2000 [5] | AUC↑     | 98.39           | 98.57           | 98.89           | 98.95           |
|         | F1↑      | 84.94           | 86.59           | 85.79           | 87.26           |
|         | SSIM↑    | 0.22            | 0.43            | 0.25            | 0.39            |
|         | DAC↑     | 1.05            | 1.78            | 1.33            | 2.01            |

**p < 0.001, * p < 0.01, * p < 0.025, ns: p > 0.025**

F. Comparisons With Prominent Methods

In this section, we compare our SGT model with other prominent baseline methods and attention-guidance methods on INbreast [52] and FIGRIM [5] datasets, such as the different variations of ResNet [24], U-Net [59], EfficientNet-B0 [68], EfficientNet-B7 [68], and prominent method Swin transformer [44]. We also compare SIBNet [48], which uses a different approach to alleviate the shortcut learning problem. As shown in the three to nine rows in Table V, our proposed SGT-S model significantly outperforms the other baseline methods in F1 and SSIM score on INbreast [52], except for AUC that is slightly inferior to ResNet-101. However, when we increase the complexity of the model, i.e., when using SGT-B, the model takes the optimal results on all four metrics. In the FIGRIM [5] dataset, our proposed SGT model is more advanced in both AUC and F1 but not as good as ResNet in SSIM and DAC metrics.

In the medical image datasets, we believe that our model successfully learns the radiologists’ prior knowledge and effectively guides the training of the model to achieve competitive results. It can also be seen in Table V that unguided baseline models are more likely to learn shortcut knowledge in medical diagnosis, which will also have more serious consequences. In the natural image datasets, we believe that what causes SSIM and DAC to be inferior to CNN architecture models is the inductive bias inherent in CNN models. This inductive bias is more consistent with the human visual system and more common even with a small amount of training data in natural images, which is an advantage of the CNN architecture models, while the ViT models have difficulty learning this inductive bias from small datasets. Thus, the attention-related metrics (SSIM and DAC) of the CNN models in Table V are closer to those of humans. However, the transformer-based model is superior in its ability to extract image information, so it can be seen that, after effective guidance, our SGT model outperforms the CNN architecture models in both AUC and F1 metrics.
TABLE V
COMPARISON RESULTS WITH OTHER PROMINENT METHODS ON INBREAST AND FIGRIM DATASETS. THE AUC, F1, SSIM, AND DAC ARE REPORTED.
**BOLD** AND **UNDERLINE** DENOTE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

| Method                  | Params | FLOPs | INBREAST [52] | FIGRIM [5] |
|-------------------------|--------|-------|---------------|------------|
|                         |        |       | AUC†          | F1†        | SSIM†       | DAC†        | AUC†          | F1†        | SSIM†       | DAC†        |
| ViT-SC (Baseline) [15]  | 22M    |       | 88.79         | 91.01      | 0.47***     | 1.81        | 99.47         | 95.12      | 0.33***     | 0.24        |
| ViTB (Baseline) [15]    | 86M    |       | 90.18         | 90.54      | 0.49***     | 2.16        | 99.77         | 94.94      | 0.30***     | 0.26        |
| ResNet-18 [24]          | 11M    |       | 98.86         | 88.55      | 0.51**      | 3.15        | 99.44         | 86.68      | 0.64***     | 0.36        |
| ResNet-50 [24]          | 24M    |       | 93.55         | 89.35      | 0.53***     | 4.60        | 99.57         | 89.55      | 0.64**      | 0.38        |
| ResNet-101 [24]         | 43M    |       | 94.00         | 90.71      | 0.58***     | 7.78        | 99.57         | 90.09      | 0.67**      | 0.41        |
| U-Net [59]              | 6M     |       | 91.48         | 87.27      | 0.49***     | 1.77        | 97.80         | 83.96      | 0.47***     | 0.29        |
| EfficientNet-B0 [68]    | 4M     |       | 88.84         | 87.20      | 0.50***     | 2.44        | 98.64         | 84.18      | 0.59*       | 0.27        |
| EfficientNet-B7 [68]    | 6M     |       | 92.08         | 91.64      | 0.55**      | 3.91        | 99.47         | 90.66      | 0.63***     | 0.29        |
| SwinT Vi [44]           | 49M    |       | 88.84         | 91.64      | 0.48***     | 4.40        | 99.32         | 92.75      | 0.25**      | 0.24        |
| SIBNet [48]             | 11M    |       | 82.77         | 79.63      | 0.39***     | 1.56        | 98.16         | 85.27      | 0.36**      | 0.22        |

| Method          | Params | FLOPs | INBREAST [52] | FIGRIM [5] |
|-----------------|--------|-------|---------------|------------|
| ResNet-18+Gaze [71] | 11M    |       | 95.37         | 88.96      | 0.56*       | 4.39        | 99.53         | 87.75      | 0.62***     | 0.33        |
| ResNet-50+Gaze [73] | 24M    |       | 91.22         | 87.59      | 0.48***     | 4.27        | 99.62         | 88.26      | 0.60*       | 0.32        |
| ResNet-101+Gaze [73] | 43M    |       | 93.65         | 90.89      | 0.54**      | 5.32        | 99.69         | 91.22      | 0.65**      | 0.35        |
| U-Net+Gaze [31]      | 6M     |       | 90.77         | 86.37      | 0.48***     | 2.63        | 98.99         | 89.60      | 0.57**      | 0.29        |

| Method          | Params | FLOPs | INBREAST [52] | FIGRIM [5] |
|-----------------|--------|-------|---------------|------------|
| SGT-S (ours)    | 22M    |       | 93.63         | 91.77      | 0.74        | 7.80        | 99.84         | 94.84      | 0.61        | 0.34        |
| SGT-B (ours)    | 86M    |       | 96.91         | 92.71      | 0.77        | 8.94        | 99.81         | 95.24      | 0.58        | 0.36        |

* * * : p < 0.001, ** : p < 0.01, * : p < 0.025, ns : p > 0.025, – : statistic < 0

We also compare the human-attention-guidance methods, ResNet + Gaze [73] and U-Net + Gaze [31], both of which achieve guidance by directly participating in the model loss calculation through the attention heatmap generated by the human eye-gaze. As shown in the middle part of the Table V, the improvement by directly involving human attention in the loss calculation is limited and even shows a downward trend on U-Net + Gaze [31]. Our approach further exploits the advantages of the transformer architecture to improve the interpretability of the model while enhancing performance.

V. LIMITATION
In this work, we used the saliency map predicted by the saliency prediction model to guide the ViT model and rectify the shortcut learning problem. However, the predicted saliency map itself may have a shortcut learning problem. In fact, the loss function of the saliency prediction task differs from the classification task in which it uses a stronger 2-D eye-tracking heatmap or 2-D mask as the ground truth for calculating the loss. Thus, in the training phase, learning saliency positions hardly induce shortcut learning problems. We believe that the shortcut problem of the saliency map mentioned above is actually the problem of labeled training data, i.e., eye-gaze data. Also, as mentioned in Section III-D, collected eye-gaze data sometimes have a small amount of human bias, which causes the saliency prediction model to misguide the SGT model. In this work, we obtain generic human attention by training the saliency model to minimize the error from individuals. We use a simple random guidance strategy to mitigate potential errors caused by bias. However, how to effectively balance human prior knowledge and the model’s own learning still needs to be explored.

VI. CONCLUSION AND DISCUSSION
In this work, we revealed that the shortcut learning phenomenon is prevalent in the ViT framework through SSIM and self-defined DAC metrics, as well as visual analysis of the model’s attention. We proposed a novel SGT to suppress and rectify shortcut learning by infusing artificial prior knowledge. This artificial prior knowledge comes from the saliency map predicted by a saliency model trained on real human eye-gaze data, which allows us to use the SGT model on images without eye-gaze data. We explored a simple strategy to balance the guidance of human prior knowledge with learning the model itself. In our experiments, we found that the guidance of human prior knowledge works better on medical images than on natural images. It proves that human prior knowledge, especially in specialized fields, is of great potential value.

In the model architecture, the mask encoder that we use can be applied to a variety of transformer architectures with great flexibility. For the saliency prediction module, we believe that the incorporation of additional domain knowledge, such as cognitive science [21] and brain function [30], will further enhance the interpretability of the models and yield prediction outcomes that are more aligned with human cognition. For the model design, we believe that multitask learning may suppress certain shortcut learning brought by a single task, and task-driven eye-gaze data in natural images could provide more significant improvement. In the future, we will do further research based on this and design a more comprehensive evaluation framework for the shortcut learning phenomenon.

REFERENCES
[1] A. Borji, “Saliency prediction in the deep learning era: Successes and limitations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 2, pp. 679–700, Feb. 2021.
[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, “Salient object detection: A survey,” Comput. Vis. Media, vol. 5, no. 2, pp. 117–150, Jun. 2019.
[3] A. Borji and L. Itti, “CAT2000: A large scale fixation dataset for boosting saliency research,” 2015, arXiv:1505.03581.
[4] A. Brown, N. Tomasev, J. Freyberg, Y. Liu, A. Karthikesalingam, and J. Schrouff, “Detecting shortcut learning for fair medical AI using shortcut testing,” 2022, arXiv:2207.10384.
[5] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, “Intrinsic and extrinsic effects on image memorability,” *Vis. Res.*, vol. 116, pp. 165–178, Nov. 2015.

[6] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What do different evaluation metrics tell us about saliency models?” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[7] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.

[8] Y. Chen et al., “A unified and biologically-plausible relational graph representation of vision transformers,” 2022, arXiv:2206.11073.

[9] Y. Chen et al., “Mask-guided vision transformer (MG-ViT) for few-shot learning,” 2022, arXiv:2205.09995.

[10] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multi-level network for saliency prediction,” in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.

[11] P. G. Correia and H. L. Cardoso, “Towards explaining shortcut learning through attention visualization and adversarial attacks,” in *Proc. Int. Conf. Eng. Appl. Neural Netw. Cham, Switzerland*: Springer, 2023, pp. 558–569.

[12] H. Dai et al., “SAMAug: Point prompt augmentation for segment anything model,” 2023, arXiv:2307.01187.

[13] C. Danzette, R. Cadène, D. Teney, and M. Cord, “Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1554–1563.

[14] 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.

[15] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale” in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[16] R. Droste, J. Jiao, and J. A. Noble, “Unified image and video saliency modeling,” in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland*: Springer, 2020, pp. 419–435.

[17] M. Du, F. He, N. Zou, D. Tao, and X. Hu, “Shortcut learning of large language models in natural language understanding,” 2022, arXiv:2208.11857.

[18] 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.

[19] A. T. Duchowski and A. T. Duchowski, “Eye Tracking Methodology. Cham, Switzerland: Springer, 2017.

[20] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, “JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3049–3059.

[21] J. Gao et al., “Prediction of cognitive scores by joint use of movie-watching fMRI connectivity and eye tracking via attention-CensNet,” in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6412–6421.

[22] N. Khosravan et al., “Gaze2Segment: A pilot study for integrating eye-tracking technology into medical image segmentation,” in *Bayesian and graphical Models for Biomedical Imaging Cham, Switzerland*: Springer, 2016, pp. 94–104.

[23] S. Mall, E. Krupinski, and C. Mello-Thoms, “Missed cancer and visual function, cognition, and behavior,” *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 103502.

[24] T. Niven and H.-Y. Kao, “Probing neural network comprehension of natural language arguments,” in *Proc. 57th Ann. Meet. Assoc. Comput. Linguistics*, 2019, pp. 3428–3448.

[25] I. Moreira et al., “INBreast: Toward a full-field digital mammographic database,” *Academic Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.

[26] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[27] 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.

[28] 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, no. 3, Aug. 2018, Art. no. 035052.

[29] S. Mall, E. 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,” *Proc. SPIE*, vol. 10952, pp. 281–287, Mar. 2019.

[30] T. McCoy, E. Pavlick, and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference,” in *Proc. 57th Ann. Meet. Assoc. Comput. Linguistics*, 2019, pp. 3428–3448.
[55] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 8748–8763.

[56] A. Trivedi et al., “Deep learning models for COVID-19 chest X-ray classification: Preventing shortcut learning using feature disentanglement,” PLoS ONE, vol. 17, no. 10, p. e0274098, 2022.

[57] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra, “Can contrastive learning avoid shortcut solutions?,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 4974–4986.

[58] Y. Rong, W. Xu, Z. Akata, and E. Kasneci, “Human attention in fine-grained classification,” 2021, arXiv:2111.01628.

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

[60] K. Saab et al., “Observational supervision for medical image classification using gaze data,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2021, pp. 603–614.

[61] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2. Inverted residuals and linear bottlelenecks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4510–4520.

[62] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, “Gaze-based interaction for semi-automatic photo cropping,” in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2006, pp. 771–780.

[63] 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.

[64] Q. Si et al., “Language prior is not the only shortcut: A benchmark for shortcut learning in VQA,” 2022, arXiv:2210.04692.

[65] (2020). SIIM-ACR Pneumothorax Segmentation. [Online]. Available: https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

[66] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1536.

[67] J. N. Stemmer et al., “Eye tracking for deep learning segmentation using convolutional neural networks,” J. Digit. Imag., vol. 32, no. 4, pp. 597–604, Aug. 2019.

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

[69] R. Tang, D. Kong, L. Huang, and H. Xue, “Large language models can be lazy learners: Analyze shortcuts in in-context learning,” 2023, arXiv:2305.17256.

[70] H. Touvron et al., “Training data-efficient image transformers & distillation through attention,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 10347–10357.

[71] S. Valipour, M. Siam, M. Jagersand, and N. Ray, “Recurrent fully convolutional networks for video segmentation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2017, pp. 29–36.

[72] J. Wang et al., “Review of large vision models and visual prompt engineering,” 2023, arXiv:2307.00855.

[73] 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.

[74] W. Wang, J. Shen, X. Dong, and A. Borji, “Salient object detection driven by fixation prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1711–1720.

[75] 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.

[76] Z. Wu, L. Su, and Q. Huang, “Stacked cross refinement network for edge-aware salient object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7264–7273.

[77] Z. Xiao et al., “Instruction-VIT: Multi-modal prompts for instruction learning in VIT,” 2023, arXiv:2305.00201.

[78] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 3166–3173.

[79] 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.
Tuo Zhang received the B.E. and Ph.D. degrees from Northwestern Polytechnical University, Xi’an, China, in 2007 and 2015, respectively. He is currently a Research Associate with Northwestern Polytechnical University. His research interests include machine learning and medical image analysis.

Xintao Hu received the Ph.D. degree from Northwestern Polytechnical University (NWPU), Xi’an, China, in 2011. He is currently a Professor with the School of Automation, NWPU. His research interests include brain image computing and brain-inspired artificial intelligence.

Dinggang Shen (Fellow, IEEE) was the Director of Faculty Development of the Department of Radiology, the Director of the Imaging Information Center, the Director of the IDEA Lab, and the Director of Image Analysis Core of the Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC, USA. He is currently the Founding Dean of the School of Biomedical Engineering, ShanghaiTech University, Shanghai, China. Before joining ShanghaiTech University, he was a tenured Professor of radiology, computer science, and biomedical engineering with BRIC, UNC. He is also the Co-CEO of the Department of Research and Development, Shanghai United Imaging Intelligence Company Ltd., Shanghai. He has published more than 1100 papers, with an H-index of 117 and more than 55000 citations. He has been involved in the application of machine learning and artificial intelligence (AI) in medical image computing for a long time, including early brain development, early diagnosis, and the prediction of Alzheimer’s disease, as well as diagnosis, prognosis and radiotherapy of brain tumor, breast cancer, and prostate cancer. He is a pioneering scientist carrying out imaging AI research all over the world and was one of the first to apply deep learning to medical imaging in 2012.

Dr. Shen is a fellow of American Institute for Medical and Biological Engineering (AIMBE), International Association for Pattern Recognition (IAPR), and International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). He has served as the General Chair of MICCAI 2019. He serves on the editorial boards of eight international journals.

Xi Jiang received the B.E. degree in automation from Northwestern Polytechnical University, Xi’an, China, in 2009, and the Ph.D. degree in computer science from the University of Georgia, Athens, GA, USA, in 2016. Since 2017, he has been an Associate Professor with the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine-learning/deep-learning-based medical image analysis.

Dr. Jiang was a recipient of the Li Foundation Heritage Prize, USA, for outstanding research and contributions in the interdisciplinary field of brain science in 2019.

Tianming Liu (Senior Member, IEEE) is currently a Distinguished Research Professor of computer science with the University of Georgia (UGA), Athens, GA, USA. His research interests are brain imaging, computational neuroscience, and brain-inspired artificial intelligence. He has published over 400 papers in these areas. His Google Scholar citations are over 10000+. His H-index is 54.

Dr. Liu is an elected fellow of the American Institute for Medical and Biological Engineering. He was a recipient of the NIH Career Award and the NSF CAREER Award. He serves on the editorial boards of multiple journals, including Medical Image Analysis, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE REVIEWS IN BIOMEDICAL ENGINEERING, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS.