STRUCTURAL VISUAL GUIDANCE ATTENTION NETWORKS IN RETINOPATHY OF PREMATURITY

V. Yildiz[⋆], *S. Ioannidis*[⋆], *I. Yildiz*[⋆], *P. Tian*[⋆], *J. P. Campbell*[§], *S. Ostmo*[§],
J. Kalpathy-Cramer[†], *M. F. Chiang*[§], *D. Erdoğmuş*[⋆], *J. Dy*[⋆]
[⋆]Cognitive Systems Laboratory, ECE Department, Northeastern University, MA, USA
[†] Department of Radiology, Massachusetts General Hospital, MA, USA
[§] Department of Ophthalmology, Oregon Health & Science University, OR, USA

ABSTRACT

Convolutional neural networks (CNNs) have shown great performance in medical diagnostic applications. However, because their black-box nature, clinicians are reluctant to trust CNN diagnostic outcomes. Incorporating visual attention capabilities in CNNs enhances interpretability by highlighting regions in the images that CNNs utilize for prediction. Clinicians can often provide domain knowledge on relevant features: e.g., to diagnose retinopathy of prematurity (ROP), structural information such as tortuosity of vessels aid clinicians in diagnosing ROP. We propose a Structural Visual Guidance Attention Networks (SVGA-Net) method, that leverages structural domain knowledge to guide visual attention in CNNs. Experiments on a dataset of 5512 posterior retinal images, taken using a wide-angle fundus camera, show that SVGA-Net achieves 0.987 and 0.979 AUC to predict plus and normal categories, respectively. SVGA-Net consistently results in higher AUC compared to visual attention CNNs without guidance, baseline CNNs, and CNNs with structured masks.

Index Terms— Interpretability, CNN, ROP, Attention

1 Introduction

Retinopathy of prematurity (ROP) is a disease that affects premature infants and is a leading cause of childhood blindness [1]. Around 14,000-16,000 premature born infants are affected by of ROP in the U.S. each year [2]. Correct classification of three levels of ROP, *normal*, *pre-plus*, and *plus*, plays an important role for treatment planning. If infants with severe ROP are not treated promptly, the disease can lead to impaired vision or blindness [2]. As the survival rate of prematurely born babies increases, the number of infants at risk of ROP also increases [3]. Lack of access to ROP experts remains a challenge. These factors bring about a need for a trustworthy ROP detection system.

Neural networks have made a high impact on many medical tasks [4–7], including detection of ROP from fundus images [8,9]. State-of-the-art ROP detection systems employ convolutional neural networks (CNNs) [8] and achieve up to 0.947 and 0.982 area under the ROC curve (AUC) in the discrimination of *normal* and *plus* levels of ROP, respectively. However, because of their black-box nature, clinicians are reluctant to trust diagnostic outcomes from CNNs. In contrast, earlier structural feature based methods [10,11], extracting features dedicated to diagnosis of ROP achieve a lower prediction performance than CNNs, but are more interpretable.

In this paper, we aim to address this interpretability problem of CNNs. One approach to interpretability is to identify regions of an image that the network focuses on when making a prediction. Two ways of finding focus regions are (a) saliency maps, and (b) attention networks. Saliency maps identify focus regions by looking at layer outputs of a trained network[12–16]. Recent studies have shown that saliency maps tend to bias towards high magnitude inputs, rather than truly discriminative areas [17]. Attention networks identify the focus regions of a CNN by explicitly incorporating attention in the network architecture via the generation of soft masks, that weigh the outputs of intermediate layers. Attention masks in CNNs can be learned to weigh the spatial information [18,19] or the channels of layer outputs [20,21].

When used for spatial attention, attention maps suppress regions irrelevant for target classification. This idea aligns well with an actual diagnosis process, as clinicians only focus on relevant regions while diagnosing a disease. However, standard attention architectures try to learn attention regions via class label supervision alone. In this paper, we propose a method improving attention maps with additional domain guided information, mimicking (and incorporating expertise from) clinician diagnostic decision-making.

Clinicians' focus on relevant regions is based on their knowledge of the disease. For example, when diagnosing ROP, ophthalmologists focus on the posterior retinal blood vessels with abnormal tortuosity and dilation [22]. We believe that such structural domain knowledge can be leveraged to enhance CNN performance and aid interpretability. Most of the current attention architectures do not provide any guidance on attention regions of networks. Several studies [23–25] guide the network attention via manually annotated segmentation of classification target; however, in contrast to our proposed approach, they do not incorporate structural domain knowledge.

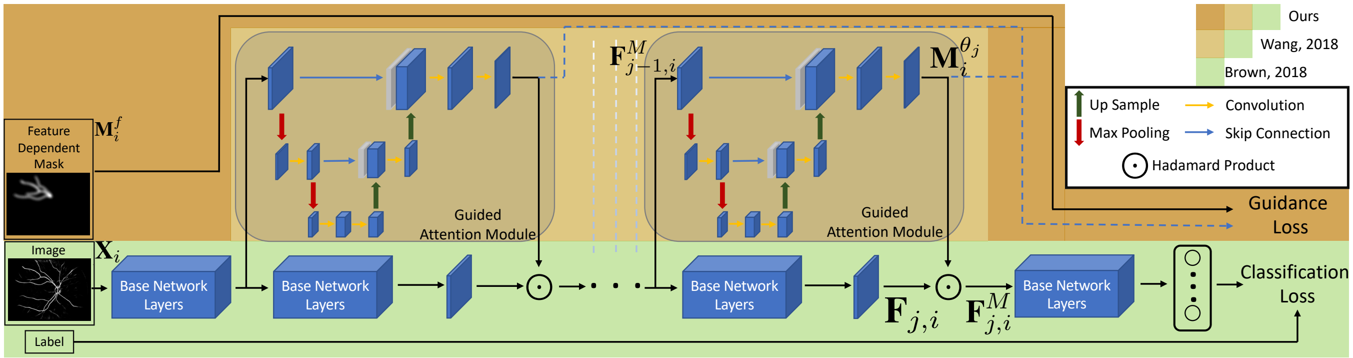

Fig. 1: Proposed Structural Visual Guidance Attention Networks, SVGA-Net.

In this paper, we propose structural visual guidance attention networks (SVGA-Net). Unlike standard visual attention networks where the network is required to learn attention itself, or supervised attention networks where manual segmentations are needed, we guide the attention of the network to the regions that are relevant to clinicians. We do so by presenting a structural way to generate domain knowledge based attention masks for ROP diagnosis and by guiding the network attention to these masks. Our contributions are twofold. First, we achieve state-of-the-art results in automated diagnosis of ROP (0.979 and 0.987 AUC detecting *normal* and *plus* levels, respectively). Second, we make ROP detection network more interpretable by highlighting the regions where network focuses on when making a prediction, without degrading performance.

2 Method

A convolutional layer in a CNN convolves its entire input with multiple filters to generate feature maps at the output of the layer. Often, the target class appears in a small region of the image. Thus, using the information from the entire image may increase the noise level in the feature maps because of the features extracted outside of the target class region. One way to overcome this problem is to use *attention* which highlights regions that are important for classification and suppresses other regions. Attention methods [19, 20, 23, 25] generate an attention mask applied to feature maps in the network. Such masks also improve the interpretability of CNNs as the network can answer the question of which regions in the image it focuses on when making a prediction.

Given a dataset containing N images, indexed by $i \in \{1, 2, \ldots, N\}$, every image i is represented as $\mathbf{X}_i \in \mathbb{R}^{h \times w}$ where h and w are the height and width of the image, respectively. For each image $\mathbf{X}_i$, a labeler generates a label $y_i \in \{$*normal, pre-plus, plus*$\}$, which indicates the ROP level, with *plus* being the most severe. Let mask $\mathbf{M}_i^f \in [0, 1]^{h \times w}$ be a domain knowledge guided mask. **Our goal is to learn CNNs that perform classification well, and simultaneously learn attention masks that are more interpretable as guided by domain knowledge masks $\mathbf{M}_i^f$.**

Wang et al. [18] construct visual attention networks by attaching attention modules on a base network. The parts highlighted in green and light brown in Fig. 1 present the structure of visual attention networks. We choose this architecture because of the flexibility in the base architecture and modular design of attention modules and base networks.

As shown in Fig. 1, attention modules are effectively additional layers generating masks. Let $j \in \{1, .., n\}$ be the index of the j-th attention module parameterized by θ_j, and $\mathbf{F}_{j,i}$, $\mathbf{F}_{j,i}^M \in \mathbb{R}^{c \times h_j \times w_j}$ be the feature maps of $\mathbf{X}_i$ before and after j-th attention mask is applied, respectively. Feature maps $\mathbf{F}_{j-1,i}^M$ and attention mask $\mathbf{M}_i^{\theta_j} \in [0, 1]^{h_j \times w_j}$ are the input-output pair of the j-th attention module. For every $c \in \{1, ..., C\}$, which is the index of the feature map channel, attention masks alter the feature maps $\mathbf{F}_{j,i}$ via: $\mathbf{F}_{j,i,c}^M = \mathbf{F}_{j,i,c} \odot \mathbf{M}_i^{\theta_j}$, where $\odot$ is the Hadamard product.

Our attention module uses UNet-like bottom-up top-down structure with skip connections [4]. This architecture is similar to Wang et al. [18], with the difference that we use convolutional units instead of residual units.

In baseline visual attention networks, attention module and base network layers are trained jointly by minimizing the categorical cross entropy loss for classification $\mathcal{L}_C(\mathbf{X}_i, y_i)$.

2.1 Structural Visual Guidance Attention Networks: SVGA-Net

To improve classification performance as well as provide interpretability, our loss function consists of two terms: (a) classification loss $\mathcal{L}_C(\mathbf{X}_i, y_i)$ and (b) guidance loss $\mathcal{L}_G(\mathbf{X}_i, \mathbf{M}_i^f)$. We use categorical cross entropy loss for $\mathcal{L}_C(\mathbf{X}_i, y_i)$. Given image $\mathbf{X}_i$, we generate a mask that contains clinically important regions $\mathbf{M}_i^f \in [0, 1]^{h \times w}$, as described in Section 2.2. We use these to guide the attention masks generated by the network $\mathbf{M}_i^{\theta_j} \in [0, 1]^{h_j \times w_j}$, via the guidance loss, we define as:

$$\mathcal{L}_G(\mathbf{X}_i, \mathbf{M}_i^f) = \sum_j \frac{1}{h_j w_j} \|\mathbf{M}_i^f - \mathbf{M}_i^{\theta_j}\|_F^2, \tag{1}$$

where the Frobenius matrix norm $\|\cdot\|_F : \mathbb{R}^{h_j \times w_j} \to \mathbb{R}$ is used.

Our proposed SVGA-Net loss function is defined as:

$$\mathcal{L}(\mathbf{X}_i, y_i, \mathbf{M}_i^f) = \mathcal{L}_C(\mathbf{X}_i, y_i) + \lambda \mathcal{L}_G(\mathbf{X}_i, \mathbf{M}_i^f), \tag{2}$$

where λ is a trade-off control parameter between classification and guidance loss.

2.2 Structural Feature Dependent Attention Mask Generation

When diagnosing ROP, clinicians focus on highly tortuous and dilated blood vessels [22]. Following the pipeline from Yildiz et al. [10], we compute *Cumulative Tortuosity Index (CTI)* and *Average Segment Diameter (ASD)* measuring tortuosity and dilation to generate feature dependent masks.

For every image $\mathbf{X}_i$, we generate two masks, $\mathbf{M}^f_{i,CTI} \in [0,1]^{h \times w}$ and $\mathbf{M}^f_{i,ASD} \in [0,1]^{h \times w}$, around the 20% of vessel segments that have the highest *CTI* and *ASD*. We also generate a mask $\mathbf{M}^f_i \in [0,1]^{h \times w}$ that contains both high *CTI* and *ASD* regions by taking element-wise maximum of masks $\mathbf{M}^f_{i,CTI}$ and $\mathbf{M}^f_{i,ASD}$. We present a sample $\mathbf{M}^f_i$ in Fig. 1.

3 Experimental Evaluation

Dataset. Our dataset contains 5512 retinal fundus images. According to its disease level, clinicians assign a label (as plus, pre-plus or normal) to each image following a reference standard diagnosis [26]. The dataset contains 163 plus, 802 pre-plus, and 4547 normal images. We use images in which vessels are segmented via the procedure proposed by Brown et al.[8].

Evaluation Metrics. We binarize labels as (a) plus vs. other classes (PvO), and (b) normal vs. other classes (NvO). We calculate AUC, accuracy (ACC), F1 and Area Under the Precision-Recall Curve (PRAUC) scores with five fold cross-validation. We present the mean of five folds and calculate the 95% confidence intervals as $1.96 \times \sigma_A$, where σ_A is the standard deviation, we calculate following Hanley et al. [27].

Base CNN Architecture and Training Details. As the base network to attach attention, we employ Inception v.1 architecture [28], which has shown great performance in many classification tasks including ROP [8]. We initialize the network weights with pretrained weights on ImageNet[29]. We employ stochastic gradient descent with learning rate 0.0001 for 100 epochs to optimize the network weights.

Competing Methods. We explore the effects of guiding attention in classification of ROP in five different setups. When used, we attach an attention module to *inception 3a* block.

No Attention[8]: We train the base CNN architecture without any attention as in Brown et al.[8]'s ROP classification model.

Feature Dependent Attention Mask: We measure the discriminative power of feature dependent masks $\mathbf{M}^f_i$ by training an Inception v.1 architecture with $\mathbf{M}^f_i$ applied to feature maps at the output of the *inception 3a* block. Note that network attention here is fixed to the feature dependent masks $\mathbf{M}^f_i$.

Unguided Attention [18]: We train the baseline visual attention network as explained in Section 2.

Guided Attention (SVGA-Net): We use the feature dependent masks for guiding the network attention as explained in the Section 2.1. In addition, we explore the effect of changing hyperparameter λ in Eq. (2) by repeating the same experiment with λ values in $\{1, 10, 15, 25, 35, 50\}$.

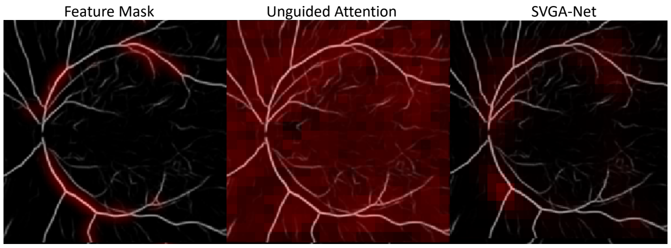

Fig. 2: Example image overlaid with its feature dependent attention mask (left), attention mask generated by unguided attention network (middle), and SVGA-Net (right). (Best printed in color.)

3.1 Results

Metric Comparison. Table 1 shows cross validation results of training to predict normal vs other and plus vs other categories. SVGA-Net achieves higher scores than other methods in predicting normal and plus categories. When we apply structural feature dependent masks to feature maps as explained in Section 3, networks achieve 0.932 and 0.960 AUC in predicting the same categories. This experiment shows that feature dependent masks indeed contain relevant information about the disease. This is because networks achieve AUCs higher than 0.93 even when forced to make predictions based on only the masked regions. Lastly, unguided attention [18] achieves almost the same AUC as no attention. This is because attention modules attached to unguided attention are only additional layers, and networks are trained only on classification loss.

Table 1: Cross validation results for competing methods

	Task		*Fixed Attention*	*Learned Attention*	
		No Attention[8]	Feature Mask	Unguided Attention [18]	Guided with Feature Mask
AUC	NvsO	0.947(0.024)	0.932(0.027)	0.942(0.025)	**0.979(0.015)**
	PvsO	0.982(0.006)	0.96(0.009)	0.981(0.006)	**0.987(0.005)**
ACC	NvsO	0.872(0.001)	0.862(0.001)	0.863(0.001)	**0.924(0.001)**
	PvsO	0.916(0.0)	0.908(0.0)	0.938(0.0)	**0.950(0.0)**
F1	NvsO	0.917(0.001)	0.911(0.001)	0.911(0.001)	**0.952(0.0)**
	PvsO	0.954(0.0)	0.95(0.0)	0.967(0.0)	**0.974(0.0)**
PR-AUC	NvsO	0.832(0.039)	0.799(0.041)	0.825(0.039)	**0.92(0.029)**
	PvsO	0.726(0.019)	0.625(0.02)	0.698(0.002)	**0.761(0.018)**

Sensitivity to λ. We train SVGA-Net with different values of λ. Fig. 3 presents the change of AUC in predicting normal and plus categories w.r.t. λ. Drastic decrease at $\lambda = 25$ is because guidance loss in Eq. (2) becomes too dominant compared to classification loss. Since guidance loss is calculated in an early layer of the network, i.e. *inception 3a*, and gradients are backpropagated from the last layer, the network does not learn classification or attention.

From left to right, Fig. 2 shows an example image blended with its feature dependent mask, attention mask generated by the baseline visual attention network [18], and attention masks generated by SVGA-Net. Red colored regions are the network focus regions when making a prediction. Without guidance, network focus is on the entire image. In contrast, as we introduce guidance, network focus regions become clearer. Compared to unguided attention [18], SVGA-Net is more interpretable, as the regions it focuses on when making a prediction

Table 2: The results of using *CTI* and *ASD* features separately.

	Task		*Fixed Attention*			*Learned Attention*		
		No Attention[8]	Feature Mask ASD Only	Feature Mask CTI Only	Feature Mask ASD and CTI	Guided with ASD Only	Guided with CTI Only	Guided with ASD and CTI
AUC	NvsO	0.947(0.024)	0.85(0.037)	0.837(0.038)	0.932(0.027)	0.942(0.025)	0.942(0.025)	**0.979(0.015)**
	PvsO	0.982(0.006)	0.914(0.012)	0.898(0.013)	0.96(0.009)	0.976(0.007)	0.976(0.007)	**0.987(0.005)**
ACC	NvsO	0.872(0.001)	0.796(0.001)	0.781(0.001)	0.862(0.001)	0.865(0.001)	0.867(0.001)	**0.924(0.001)**
	PvsO	0.916(0.0)	0.858(0.0)	0.796(0.0)	0.908(0.0)	0.918(0.0)	0.923(0.0)	**0.950(0.0)**
F1	NvsO	0.917(0.001)	0.866(0.001)	0.856(0.001)	0.911(0.001)	0.912(0.001)	0.914(0.001)	**0.952(0.0)**
	PvsO	0.954(0.0)	0.921(0.0)	0.883(0.0))	0.95(0.0)	0.956(0.0)	0.958(0.0)	**0.974(0.0)**
PR-AUC	NvsO	0.832(0.039)	0.633(0.047)	0.61(0.047))	0.799(0.041)	0.825(0.039)	0.825(0.039)	**0.92(0.029)**
	PvsO	0.726(0.019)	0.393(0.019)	0.385(0.018)	0.625(0.02)	0.679(0.02)	0.681(0.002)	**0.761(0.018)**

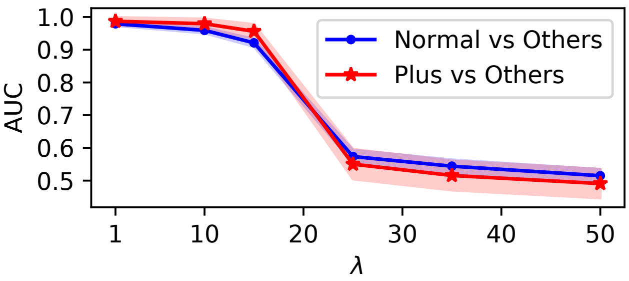

Fig. 3: Hyperparameter search

can be explained via the structural features we use.

Sparsity. Visual sparsity of feature dependent attention masks brings about the question of whether the effect of using guided feature dependent masks is due to their sparsity or the information contained in the mask. We answer this question using randomized masks in guiding the attention network. For every image $\mathbf{X}_i$, we generate a random mask $\mathbf{M}_i^r \in \{0,1\}^{h\times w}$ that contains the same number of non-zero elements as $\mathbf{M}_i^f$. We repeat *Feature Dependent Attention Mask* and *Guided Attention* experiments using $\mathbf{M}_i^r$.

Results in Table 3 show that performance improvement in SVGA-Net is not only because the feature dependent attention masks are sparse. The networks consistently achieve lower performance metrics when random mask $\mathbf{M}_i^r$ is used instead of $\mathbf{M}_i^f$. The results show that using domain knowledge in guiding attention maps improves performance more than the effect of sparsity. In fact, the fact that SVGA-Net guided with random masks achieving almost the same performance as unguided attention network [18] suggests that incorporating domain knowledge in guidance improves the network performance.

Table 3: The effect of sparsity in feature masks.

	Task	*Fixed Attention*		*Learned Attention*		
		Random Mask	Feature Mask	Unguided Attention [18]	Guided with Random Mask	Guided with Feature Mask
AUC	NvsO	0.893(0.033)	0.932(0.027)	0.942(0.025)	0.945(0.025)	**0.979(0.015)**
	PvsO	0.968(0.008)	0.96(0.009)	0.981(0.006)	0.979(0.006)	**0.987(0.005)**
ACC	NvsO	0.822(0.001)	0.862(0.001)	0.863(0.001)	0.876(0.001)	**0.924(0.001)**
	PvsO	0.885(0.0)	0.908(0.0)	0.938(0.0)	0.936(0.0)	**0.950(0.0)**
F1	NvsO	0.883(0.001	0.911(0.001))	0.911(0.001)	0.92(0.001)	**0.952(0.0)**
	PvsO	0.937(0.0)	0.95(0.0)	0.967(0.0)	0.966(0.0)	**0.974(0.0)**
PR-AUC	NvsO	0.728(0.045)	0.799(0.041)	0.825(0.039)	0.829(0.039)	**0.92(0.029)**
	PvsO	0.62(0.02)	0.625(0.02)	0.698(0.002)	0.697(0.02)	**0.761(0.018)**

Feature Analysis. To explore the discriminative power of curvature (*CTI*) and dilation (*ASD*) features separately, we repeat *Feature Dependent Attention Mask* and *Guided Attention* experiments using separate *CTI* and *ASD* dependent masks, i.e., $\mathbf{M}_{i,CTI}^f$ and $\mathbf{M}_{i,ASD}^f$. Table 2 shows that if we use only $\mathbf{M}_{i,ASD}^f$ in *Feature Dependent Attention Mask*, networks achieve 0.850 and 0.914 AUC when predicting normal vs others and plus vs others, respectively. Whereas, using only $\mathbf{M}_{i,CTI}^f$ results in 0.837 and 0.898 AUC when predicting the same categories. The AUC drops more when $\mathbf{M}_{i,CTI}^f$ is used as fixed attention at feature dependent attention masks experiment compared to $\mathbf{M}_{i,ASD}^f$. The results suggest that highly dilated vessels are more discriminative for both normal vs others and plus vs others categories than highly tortuous vessels.

4 Conclusion

In this paper, we propose a principled way to improve interpretability of CNNs. We guide network attention to image regions that contain clinically relevant information. We improve the interpretability of the ROP classification network without degrading its performance. We have introduced a novel way for incorporating domain knowledge in learning attention maps in CNNs. It would be interesting to apply SVGA-Net to other medical diagnosis applications and explore if like in our ROP case, domain guided attention masks can lead to significant improvements.

5 Compliance with Ethical Standards

This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University) and at each of eight study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center and Asociacion para Evitar la Ceguera en Mexico). This study was conducted in accordance with the Declaration of Helsinki. Written informed consent for the study was obtained from parents of all infants enrolled.

6 Acknowledgments

This work is supported by NIH (R01EY019474), NSF (SCH-1622542 at MGH; SCH-1622536 at Northeastern; SCH-1622679 at OHSU), and by unrestricted departmental funding from Research to Prevent Blindness (OHSU).

Disclosure: V.M. Yildiz, None; P. Tian, None; I. Yildiz, None; J. Kalpathy-Cramer, None; J. Dy, None; S. Ioannidis, None; D. Erdogmus, None; S. Ostmo, None; J.P. Campbell, None; M.F. Chiang, Clarity Medical Systems Scientific Advisory Board (S), Novartis (C), and Inteleretina (I)

7 References

[1] C Gilbert and A Foster, "Childhood blindness in the context of vision 2020: the right to sight," *Bulletin of the WHO*, vol. 79, 2001.

[2] National Eye Institute, "Retinopathy of prematurity," https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/retinopathy-prematurity, 2019.

[3] C Gilbert, A Fielder, L Gordillo, G Quinn, R Semiglia, P Visintin, A Zin, et al., "Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs," *Pediatrics*, vol. 115, no. 5, 2005.

[4] O Ronneberger, P Fischer, and T Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015.

[5] P Rajpurkar, J Irvin, K Zhu, B Yang, H Mehta, T Duan, D Ding, A Bagul, C Langlotz, K Shpanskaya, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[6] A Esteva, B Kuprel, R A Novoa, J Ko, S M Swetter, H M Blau, and S Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, 2017.

[7] L Hou, D Samaras, T M Kurc, Y Gao, J E Davis, and J H Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *CVPR*, 2016.

[8] J M Brown, J P Campbell, A Beers, K Chang, S Ostmo, RV P Chan, J Dy, D Erdogmus, S Ioannidis, J Kalpathy-Cramer, et al., "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," *JAMA ophthalmology*, vol. 136, no. 7, 2018.

[9] D E Worrall, C M Wilson, and G J Brostow, "Automated retinopathy of prematurity case detection with convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016.

[10] V M Yildiz, P Tian, I Yildiz, J M Brown, J Kalpathy-Cramer, J Dy, S Ioannidis, D Erdogmus, S Ostmo, S J Kim, et al., "Plus disease in retinopathy of prematurity: Convolutional neural network performance using a combined neural network and feature extraction approach," *TVST*, vol. 9, no. 2, 2020.

[11] E Ataer-Cansizoglu, V Bolon-Canedo, J P Campbell, A Bozkurt, D Erdogmus, J Kalpathy-Cramer, S Patel, K Jonas, RV P Chan, S Ostmo, et al., "Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-rop" system and image features associated with expert diagnosis," *TVST*, vol. 4, no. 6, 2015.

[12] B Zhou, A Khosla, A Lapedriza, A Oliva, and A Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[13] K Simonyan, A Vedaldi, and A Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[14] J T Springenberg, A Dosovitskiy, T Brox, and M Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.

[15] A Shrikumar, P Greenside, A Shcherbina, and A Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[16] 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 *ICCV*, 2017.

[17] J Adebayo, J Gilmer, M Muelly, I Goodfellow, M Hardt, and B Kim, "Sanity checks for saliency maps," in *NeurIPS*, 2018.

[18] F Wang, M Jiang, C Qian, S Yang, C Li, H Zhang, X Wang, and X Tang, "Residual attention network for image classification," in *CVPR*, 2017.

[19] Q Guan, Y Huang, Z Zhong, Z Zheng, L Zheng, and Y Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[20] J Hu, L Shen, and G Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[21] X Yao, D She, S Zhao, J Liang, Y Lai, and J Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *ICCV*, 2019.

[22] International Committee for the Classification of Retinopathy of Prematurity et al., "The international classification of retinopathy of prematurity revisited.," *Archives of ophthalmology (Chicago, Ill.: 1960)*, vol. 123, no. 7, 2005.

[23] K Li, Z Wu, K Peng, J Ernst, and Y Fu, "Tell me where to look: Guided attention inference network," in *CVPR*, 2018.

[24] J Son, W Bae, S Kim, S Park, and K Jung, "Classification of findings with localized lesions in fundoscopic images using a regionally guided cnn," in *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer, 2018.

[25] H Yang, J Kim, H Kim, and S P Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE TMI*, 2019.

[26] M C Ryan, S Ostmo, K Jonas, A Berrocal, K Drenser, J Horowitz, T C Lee, C Simmons, M Martinez-Castellanos, RV P Chan, et al., "Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology," in *AMIA annual symposium*. American Medical Informatics Association, 2014, vol. 2014.

[27] J A Hanley and B J McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, 1982.

[28] C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, and A Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[29] J Deng, W Dong, R Socher, L Li, K Li, and L Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009.