GazeRadar: A Gaze and Radiomics-guided Disease Localization Framework

Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna*

Stony Brook University, NY, USA {moinak.bhattacharya, shubham.jain.1, prateek.prasanna}@stonybrook.edu

Abstract. We present *GazeRadar*, a novel radiomics and eye gaze-guided deep learning architecture for disease localization in chest radiographs. *GazeRadar* combines the representation of radiologists' visual search patterns with corresponding radiomic signatures into an integrated *radiomics-visual attention representation* for downstream disease localization and classification tasks. Radiologists generally tend to focus on fine-grained disease features, while radiomics features provide high-level textural information. Our framework first 'fuses' radiomics features with visual features inside a teacher block. The visual features are learned through a teacher-focal block, while the radiomics features are learned through a teacher-global block. A novel Radiomics-Visual Attention loss is proposed to transfer knowledge from this joint radiomics-visual attention representation of the teacher network to the student network. We show that *GazeRadar* outperforms baseline approaches for disease localization and classification tasks on 4 large scale chest radiograph datasets comprising multiple diseases. [1]

Keywords: Disease localization · Eye-gaze · Fusion · Radiomics.

1 Introduction

Medical image interpretation is a complex visuo-cognitive task that requires an understanding of a disease's textural patterns and locations in the image. Previous studies have demonstrated the importance of visual search patterns, obtained from eye-gaze tracking of radiologists, in disease classification, localization [25,9,17,22,41,39] and segmentation [18]. Despite the spatially-rich information, gaze-derived attention regions do not always coincide with the actual disease regions. On the other hand, handcrafted radiomics features contain context-rich textural information that focuses on abnormalities, primarily

* corresponding author

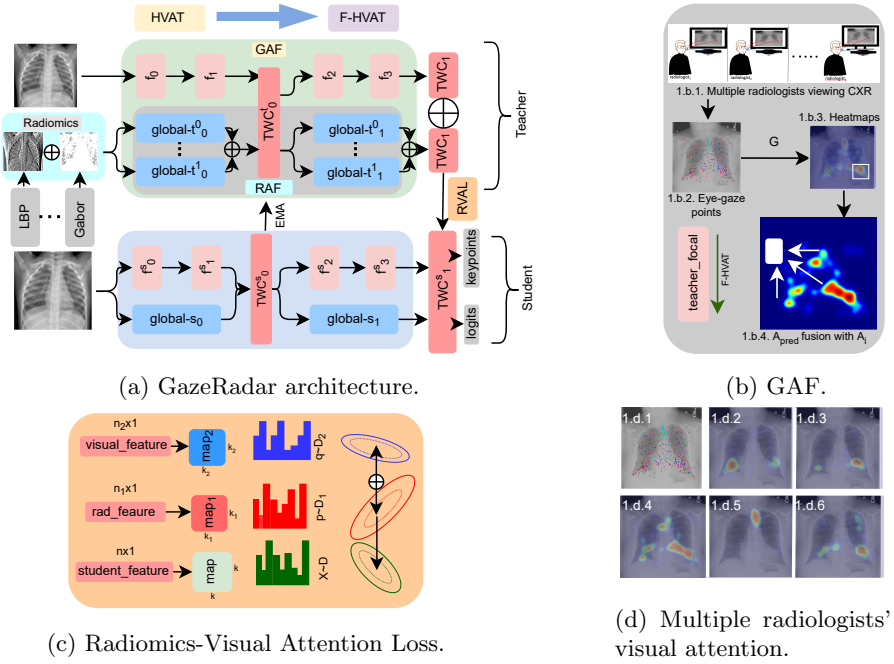
[1] Reported research was partly supported by NIH 1R21CA258493-01A1, NIH 75N92020D00021 (subcontract), and the OVPR and IEDM seed grants at Stony Brook University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Code: https://github.com/bmi-imaginelab/gazeradar

disease-specific features, manifested both within and surrounding the disease regions [30]. Several computer-aided diagnostic techniques focus on independently utilizing visual patterns and radiomics features [33,27,29]. While radiomics features have long been used for different diagnostic tasks, the concept of coupling textural features with visual attention is still unexplored.

Radiologists' visual search patterns are honed through years of training, and different levels of expertise often leads to variations in these search patterns, even on the same image [36,37]. Studies have shown that disease diagnosis can be enhanced by taking advantage of gaze patterns from multiple radiologists [2,35,16]. While Bhattacharya et al. [3] show that transformer-based architectures can leverage human visual attention from a single radiologists' gaze patterns for diagnostic tasks, they do not investigate how to fuse multiple readers' visual search patterns in a deep learning setting. Simple averaging or majority voting may lead to dispersion of attention regions or losing information regarding gaze variations.

Motivation and Overview. To address the aforementioned limitations, we propose a novel approach that couples radiomics features with visual search patterns from multiple radiologists to infer the *radiomics-visual attention*, and leverages it in a deep learning framework for improved disease detection and diagnosis. The motivation for our approach stems from a) the importance of multiple radiologists' gaze patterns in medical image interpretation, b) the importance of co-learning visual attention and textural attention, and c) distillation ability of this *radiomics-visual* attention knowledge to deep learning architectures for downstream classification and localization tasks. A global-focal learning paradigm to mimic radiologists' cognitive behavior has shown promising results in disease diagnosis [3]. This learning paradigm presents the opportunity to incorporate radiomics features as complementary attributes into the global-focal framework. Radiomics features provide a representation of disease related imaging changes as a global context, while visual attention can help characterize detailed fine-grained features into a focal context. We use this as a motivation to design a modified global-focal architecture that integrates textural information and human visual cognition with self-attention-based learning of transformers.

The main contributions of this paper are: (1) We present *GazeRadar*, a novel global-focal student-teacher architecture for disease localization based on radiomics information and visual search patterns. The teacher block learns a joint representation of radiomics and visual attention features. This representation is then used to train a student block for downstream classification and localization tasks. (2) We develop novel *Radiomics Attention Fusion* and *Gaze Attention Fusion* strategies to fuse radiomics features and gaze features, respectively. (3) We design a novel *Radiomics-Visual Attention Loss* for transferring the joint radiomics-visual knowledge from the teacher block to the student block. To the best of our knowledge, this is the first work that incorporates both, radiomics and radiologists' search patterns into a decision-making pipeline.



(a) GazeRadar architecture.

(b) GAF.

(c) Radiomics-Visual Attention Loss.

(d) Multiple radiologists' visual attention.

Fig. 1: **GazeRadar**: A radiomics-visual fusion architecture.

2 Methodology

Figure 1 presents an overview of the proposed *GazeRadar* architecture. There are two primary network blocks, the student and the teacher block. Each of these blocks consist of a global-focal network. The teacher-focal network learns human visual attention from the tracked eye gaze points of radiologists, and the teacher-global network learns attention from radiomics features. The teacher block comprises two sub-blocks, namely Gaze Attention Fusion (GAF) and Radiomics Attention Fusion (RAF). GAF module fuses visual attention regions from multiple radiologists to provide a consolidated visual attention region. This region is used for pre-training the teacher, referred to as Fused-Human Visual Attention Training (F-HVAT). Note that in the inference stage, we do not need eye-gaze data. The RAF module fuses different radiomics features using cross-attention. RAF attempts to learn 'radiomically relevant' regions from the fused radiomics features. The student network learns radiomics-visual attention from the teacher network. Henceforth, we use the following notations: g: global, f: focal, $\mathcal{D}$: probability distribution, $\mathcal{L}$: loss functions, $\mathcal{N}$: normal distribution, and the following abbreviations: GAF: Gaze Attention Fusion, RAF: Radiomics Attention Fusion, HVAT: Human Visual Attention Training, F-HVAT: Fused-HVAT, TWC: Two Way Cross, RVAL: Radiomics-Visual Attention Loss.

2.1 Teacher Block

The teacher network is designed to fuse the visual attention maps into a single representation and couple it with the radiomics attention map for downstream tasks. The teacher-focal block learns visual attention features from the eye gaze patterns of radiologists. The search patterns of different radiologists are non-uniform and hence the visual attention map may spread across different sections of the lungs, as shown in Figure 1.d.* in Figure 1. Figure 1.d.1 are eye-gaze points from different radiologists, shown in different colors. In Figure 1.d.2 - 1.d.6, the visual attention maps of 5 different radiologists are shown. The teacher-global block learns the radiomics attention.

Global-Focal network. The global and focal networks are variants of shifting window transformer architecture, inspired by [3]. In the teacher block, the global network is the RAF, and the focal network is pre-trained with GAF. The global blocks are represented as g, and the focal blocks are represented as f. There are two parallel global blocks connected with a focal block. There are k-sets of global blocks, represented as g_i^k, where $i \in \{0, 1\}$ and $k \in \{0, 1\}$, cascaded with four focal blocks represented as f_i, where $i \in \{0, 1, 2, 3\}$. Here, i is the number of shifting blocks, and k is the number of radiomics features. These blocks are connected by a Two Way Cross (TWC) module, represented as $\mathcal{C}_i(x, y)$, which is a cross-attention block between g_0^k-f_1, and g_1^k-f_3. TWC module is a cross-attention between x and y, shown as $\mathcal{C}_i(x, y) = MLP(LN_1(MHA(LN_0^0(x), LN_0^1(y)) + (x+y))) + z$, where $z = MHA(LN_0^0(x), LN_0^1(y)) + (x+y)$. Here, MHA is Multi-head attention, LN is Layer Norm and MLP is Multi-layered perceptron. The $\mathcal{C}_0(\hat{g}_0, f_1)$ is TWC_0 layer, and $\mathcal{C}_1(\hat{g}_1, f_3)$ is TWC_1 layer. Here, $\hat{g}_0 = \sum_{i=0}^{k} \lambda_i^{int} * g_0^i$, and $\hat{g}_1 = \sum_{i=0}^{k} \lambda_i^{final} * g_1^i$, where λ_k^{int} and λ_k^{final} are the intermediate and final weight parameters for k radiomics features, respectively.

Gaze Attention Fusion. The visual attention maps from n radiologists (shown in Figure 1.b.1), represented as $\mathcal{A}_i$, where $i \in \{1, 2, ..., n\}$, are first obtained as explained in Section 4 . These maps are generally localized in different sections of the raw image. The teacher block is pre-trained with single (HVAT_1 and HVAT_2, jointly termed as HVAT) and multiple (F-HVAT) radiologists' eye gaze, also shown in Figure 1.a and Figure 1.b, and discussed in Section 4. Then, the raw image is fed to this pre-trained teacher to produce a predicted visual attention map, represented as $\mathcal{A}_{pred}$. The visual attention maps from multiple radiologists, $\mathcal{A}_i$, are fused with $\mathcal{A}_{pred}$ as the reference region. This can be defined as weighted linear minimization of the distances between multiple probability distributions. The predicted probability distribution is represented as $p \sim \mathcal{N}(\mu_{pred}, \sigma_{pred}^2)$, and the probability distribution of visual attention maps from radiologists are represented as $q_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Here, $[\mu_{pred}, \sigma_{pred}]$ are the mean and standard deviation of the predicted distribution, and $[\mu_i, \sigma_i]$ are the mean and standard deviation of the distribution from visual attention regions. The n-Gaze Attention Loss (n-GAL) is the weighted distance between the p and q_i, represented as $\mathcal{L}_{n-GAL} = -\ln(d_{\mathcal{B}})$, and $d_{\mathcal{BC}} = \sum_{i=1}^{n} \alpha_i * \sqrt{p.q_i}$. Here, $d_{\mathcal{BC}}$ is a variation of Bhattacharyya coefficient [4,5], and α_i is the parameter for weighting.

Radiomics Attention Fusion. The radiomic features are obtained from the raw images, $\mathcal{I}$, represented as $\mathcal{R}_i = \mathcal{F}_i(\mathcal{I})$, where $i \in \{0, 1, ..k\}$, as shown in Figure 1.a. Here, $\mathcal{R}_i$ are the radiomics features, $\mathcal{F}_i$ are the set of radiomic filters applied to $\mathcal{I}$. The $\mathcal{R}_i$ are fed to the global networks, shown as, $\mathcal{O}_0^i = g_0^i(\mathcal{R}_i)$. This output set is fused, as $\hat{\mathcal{O}}_0 = \sum_{i=0}^{k} \lambda_i^{int} \mathcal{O}_0^i$ and fed to $\mathcal{C}_0$ layer. The output of this $\mathcal{C}_0$ layer is represented as c_0, which is then provided to the next global set to obtain the output set, $\mathcal{O}_1^i = g_1^i(c_0)$. This output set is also fused, $\hat{\mathcal{O}}_1 = \sum_{i=0}^{k} \lambda_i^{final} \mathcal{O}_1^i$, and fed to $\mathcal{C}_1$.

2.2 Student Block

Global-Focal network. The student block is a global-focal network where two global layers are stacked in-parallel with four focal layers. Similar to the teacher, the global and focal layers of the student are variants of shifting window transformers. The two global blocks are represented as g_i^s, $i \in \{0, 1\}$, and the focal blocks are represented as f_i^s, $i \in \{0, 1, 2, 3\}$. The g_0^s and f_1^s are fed into TWC_0^s in student block, represented as $c_0^s = \mathcal{C}_0^s(g_0^s(t_g(\mathcal{I})), f_1^s(f_0^s(t_f(\mathcal{I}))))$. Here $\{t_g, t_f\}$ are the augmentations of global and focal blocks respectively. The output c_0^s is fed to the subsequent global g_1^s, and focal f_2^s layers. The final output from TWC_1^s is represented as $c_1^s = \mathcal{C}_1^s(g_1^s(c_0^s), f_3^s(f_2^s(c_0^s)))$.

Training. The teacher block is updated with the student block using exponential moving average (EMA). The output of the student block is represented as c_1^s, and the output of the teacher block is represented as c_1. The EMA is represented as $\theta_{c_1} = \delta * \theta'_{c_1} + (1 - \delta) * \theta_{c_1^s}$. Here, δ is the smoothing coefficient, θ'_{c_1} is the parameter of the teacher block, $\theta_{c_1^s}$ is the parameter of the student block, and θ_{c_1} is the updated parameter of the teacher block. The downstream tasks are classification and localization. Consequently, classification heads and detection heads are appended to the output of the student block c_1^s. The classification head outputs predicted logits of shape $[\mathcal{B}, \mathcal{N}]$, where $\mathcal{B}$ is the batch size during training, and $\mathcal{N}$ is the number of classes in the datasets. The bounding box head outputs bounding boxes of shape $[\mathcal{B}, 4]$, where $\mathcal{B}$ is the batch size mentioned before, and 4 is the number of key-points of bounding boxes, in this case, $[x_{min}, y_{min}, x_{max}, y_{max}]$. A cross-entropy loss is applied for classification, shown as, $\mathcal{L}_{cls} = \sum_{i=0}^{\mathcal{B}} \hat{y}_i \log(y_i)$, where y is the predicted logit, and $\hat{y}$ is the ground-truth logit. For bounding box regression, a weighted addition of Generalised Intersection over Union (GIoU) and Mean Squared Error (MSE) loss is applied. This loss is represented as, $\mathcal{L}_{bbox} = \lambda_1^b * \mathcal{L}_{GIoU} + \lambda_2^b * \mathcal{L}_{MSE}$, where $(\lambda_1^b, \lambda_2^b)$ are the weights for adding the bounding box losses. Here, $\lambda_1^b + \lambda_2^b$ should be equal to 1. The final loss is represented as:

$$\mathcal{L} = \lambda_0^l * \mathcal{L}_{cls} + \lambda_1^l * \mathcal{L}_{bbox} + \lambda_2^l * \mathcal{L}_{rval} \tag{1}$$

where $(\lambda_0^l, \lambda_1^l, \lambda_2^l)$ are the weights for adding the individual loss components for final loss, and $\lambda_0^l + \lambda_1^l + \lambda_2^l = 1$. $\mathcal{L}_{rval}$ is the Radiomics-Visual Attention Loss (RVAL), described in the following subsection.

2.3 Radiomics-Visual Attention Loss

The teacher outputs a joint representation of radiomics, and visual attention features. The final TWC_1 layer $\mathcal{C}_1$ takes both $\hat{g}_1$ and f_3 as input. The $\hat{g}_1$ is the fused radiomics attention representation, and f_3 is the visual attention representation. The output from this final TWC_1 layer, shown as $c_1 = \mathcal{C}_1(\hat{g}_1, f_3)$, is the joint radiomics-visual attention feature representation. As shown in Figure 1.c, the radiomics attention features are represented as $p \sim \mathcal{D}_1$, and the visual attention features are represented as $q \sim \mathcal{D}_2$, where p and q are probability distributions. The joint representation of these distributions can be represented as $\mathcal{P} \sim \mathcal{D}_{12}$. From Figure 1.c, this output is represented as $\mathcal{X} \sim \mathcal{D}$. We propose a novel loss function $\mathcal{L}_{rval}$ that calculates the distance between these probability distributions, represented as:

$$\mathcal{L}_{rval} = -\ln\left(\int \sqrt{\mathcal{X}(\hat{c}_1^s) * \mathcal{P}(\hat{c}_1)}\right) \tag{2}$$

where $\hat{c}_1^s$ is the feature map obtained after post-processing the output from the TWC_1^s of the student block, and $\hat{c}_1$ is the feature map obtained after post-processing the output from the TWC_1 of the teacher block.

3 Datasets and Environment

We use 4 datasets for developing and validating our proposed techniques: 1) RSNA Pneumonia Detection Challenge dataset [32] consisting of radiographs with presence and absence of pneumonia, 2) SIIM-FISABIO-RSNA COVID-19 Detection dataset [19] for COVID-19 classification and localization, 3) NIH Chest X-rays [38], and 4) VinBigData Chest X-ray Abnormalities Detection dataset [28] comprising 14 common thorax diseases. The training, validation, and testing splits are provided in Table 1. For experimentation, we used Google Cloud Platform (GCP) with TPUs from TensorFlow Research Cloud (TRC). All experiments are in TensorFlow and Keras v2.8.0.

4 Results and Discussion

Implementation. The HVAT comprises 2 stages, namely HVAT_1 and HVAT_2. During HVAT_1, the teacher network is pre-trained on eye-gaze data from [15,10] which contains single radiologist eye-gaze points on 1083 chest x-rays from the MIMIC-CXR [14,10] dataset. For HVAT_2, the teacher network is fine-tuned on similar eye-gaze data from REFLACX [20,21,10] which contains single radiologist eye-gaze points on 2507 chest x-rays from the MIMIC-CXR [14,10] dataset. Finally, for F-HVAT, the teacher network is further finetuned on n-radiologists' eye gaze points (in this case $n = 5$) with GAF, as explained in Section 2, for 109 chest x-rays from REFLACX [20,21,10] dataset. The attention regions, shown

Table 1: **Quantitative Comparisons**. MSE (↓) and AUC (↑) are shown for RSNA (Train=21158, Val=3022, Test=6045), SIIM (4433, 633, 1266), NIH (688, 98, 196) and VBD (47539, 6791, 13582) datasets. The comparisons are shown on baseline architectures and *GazeRadar*.

Datasets	RSNA [32]		SIIM [19]		NIH [38]		VBD [28]	
Baselines	MSE	AUC	MSE	AUC	MSE	AUC	MSE	AUC
RN-R50 [24]	03.38	90.18	10.83	82.89	05.67	54.13	02.17	95.98
RN-R101 [24]	02.82	94.57	10.47	82.57	05.87	63.61	01.97	**96.65**
RN-R152 [24]	03.26	91.00	10.44	83.54	07.16	62.21	**01.91**	96.51
CN-R50 [40]	24.74	60.66	27.84	69.89	19.85	51.06	35.11	91.51
CN-R101 [40]	05.86	77.83	45.94	38.98	30.84	45.73	37.43	65.33
CN-R152 [40]	06.57	80.68	28.51	68.55	21.18	50.37	43.13	79.78
YOLOv3 [31]	**02.73**	93.12	29.64	72.29	**04.45**	56.77	07.05	83.28
YOLOv5 [13]	04.15	72.63	12.93	72.29	04.50	56.62	07.05	83.28
ViT-B16+DH [7]	04.13	80.53	13.28	72.23	05.64	57.33	11.04	89.70
ViT-B32+DH [7]	05.08	75.49	13.32	72.84	05.22	55.53	14.49	92.76
ViT-L16+DH [7]	04.88	81.92	13.34	72.39	05.50	57.10	10.73	91.68
ViT-L32+DH [7]	05.12	72.42	13.30	72.43	05.12	56.54	10.30	88.90
CCT+DH [12]	03.99	79.28	12.89	73.50	04.54	56.77	03.83	91.20
DeTr [6]	04.45	68.97	12.96	72.29	04.63	56.77	07.06	83.28
SwinT0+DH [26]	06.35	93.51	07.14	74.42	04.57	64.10	14.87	92.50
SwinT1+DH [26]	07.09	93.75	06.97	75.46	**04.45**	78.70	15.41	91.44
Ours	03.56	**96.27**	**06.56**	**99.36**	08.57	**98.68**	12.12	94.26

as heatmaps (Figure 1.b.3), are human attention based diagnostically important areas. A Gaussian filter, represented as $\mathcal{G}$, with standard deviation, $\sigma = 64$, generates the attention heatmaps. The contours from these attention heatmaps are selected with a thresholding value of $\hat{\lambda} = 140$. Then, bounding boxes are generated from the contour with the largest area, shown in Figure 1.b.3. All the images are resized to 256×256 pixels. During HVAT_i, and F-HVAT, the output of the teacher network is a $[\hat{\mathcal{B}}, 2]$ tensor of probability values (representing two classes: normal and disease) and a $[\hat{\mathcal{B}}, 4]$ tensor of keypoints. Here, $\hat{\mathcal{B}}$ is the batch size during HVAT_i and F-HVAT.

We use an Adam optimizer with a batch size of 64 for 50 epochs. The initial learning rate (LR), set to 1×10^{-2}, is scheduled with an exponential scheduler with decay steps $= 10^5$ and decay rate $= 0.2$. There is an early stopping criteria with patience $= 20$ to minimize the validation loss. *GazeRadar* follows similar training settings as the baselines. Also, in RAF, we have the radiomic features, $\mathcal{R}_i$ generated from $\mathcal{F}_i$. In our experiments, $i \in \{0, 1\}$ where $\mathcal{F}_0$ is Local Binary Pattern (LBP) [11] and $\mathcal{F}_1$ is an orthogonal Gabor filter [8], also shown in Figure 1.a.

Comparisons and Performance Metrics. *GazeRadar* is compared against standard localization architectures like variants of RetinaNet [24], CenterNet [40], YOLOv3 [31], YOLOv5 [13], and recent vision transformer architectures appended with a detection head. The vision transformer architectures used for comparisons are ViT [7], CCT [12], DeTR [6], and Swin Transformer [26]. To measure the performance of *GazeRadar* for both classification and localization tasks, we use Mean Squared Error (MSE) and Area-under-Curve (AUC). As shown in Table 1, *GazeRadar* outperforms state-of-the-art (SOTA) on 3 datasets

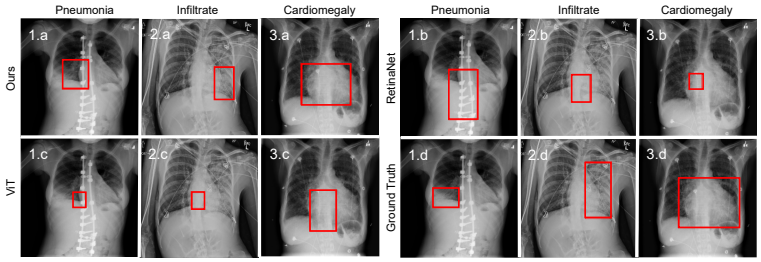


Fig. 2: **Qualitative Results**: Example localization results are shown on 3 disease types, namely Pneumonia (1.*), Infiltrate (2.*) and Cardiomegaly (3.*), for *GazeRadar* (*.a), RetinaNet (*.b) and ViT (*.c) architectures. Here, *.d are the ground-truth localizations.

(RSNA, SIIM, NIH), and achieves comparable results on VBD. We observe that *GazeRadar* outperforms the majority of vision transformer based architectures on all the datasets.

Ablation studies. In Table 2, we show the performance of different components of GAF and RAF. F-HVAT is pre-trained with HVAT_1 and HVAT_2. GAF is a component of F-HVAT. Hence, we independently evaluate HVAT_1 and HVAT_2. From Table 2, for GAF, we see that HVAT_2 performs significantly better than HVAT_1 for classification with comparable localization performance. We therefore infer that only radiomics attention, without RVAL, is not well-suited for transferring information from teacher block to student block. We also show, in the Supplementary, different components of the *GazeRadar* architecture and observe how appending different modules affects the system performance. The ablations fundamentally explain the effects of individual global-focal components, adding a teacher network, and then further training using RVAL.

Qualitative Analysis. In Figure 2, we show localization results of *GazeRadar* in comparison to RetinaNet and ViT for three different disease types: pneumonia, infiltrate, and cardiomegaly. 1.* represents Pneumonia, 2.* represents Infiltrate, and 3.* represents Cardiomegaly samples. *.a are the results from *GazeRadar*, *.b are the results from RetinaNet, *.c are the results from ViT, and *.d are the ground truth bounding boxes. We observe that the predicted bounding boxes overlap better with the ground truth for *GazeRadar* as compared with the baseline methods.

Table 2: **GAF-RAF Ablations**. MSE (↓) and AUC (↑) for NIH and SIIM datasets.

DS	NIH [38]		SIIM [19]		DS	NIH [38]		SIIM [19]	
GAF	MSE	AUC	MSE	AUC	**RAF**	MSE	AUC	MSE	AUC
HVAT_1	07.78	60.64	14.61	97.37	RAF	17.63	61.43	17.55	96.56
HVAT_2	08.39	89.77	18.60	98.38	S+RAF	08.17	59.86	23.64	80.18

5 Conclusion

This paper presents *GazeRadar*, a novel architecture that fuses radiomics and visual attention to learn a joint representation. This radiomics-visual attention is leveraged to train a student block for classification and localization tasks. A novel Radiomics-Visual Attention Loss (RVAL) is proposed to calculate the distance between the student block attention distribution and the joint representation. We demonstrated the feasibility of this approach with two radiomics features; however, as described in the methodology, this may be readily extended to other features and also to 3D imaging modalities. Our results demonstrate that radiomics and radiologists' visual search patterns harbor important complementary cues regarding disease characteristics and its location; these features can be leveraged in a deep learning framework using a student-teacher architecture. Future work will involve incorporation of RVAL in lung nodule classification [1] and computer-assisted intervention tasks [23,34].

References

1. Beig, N., Khorrami, M., Alilou, M., Prasanna, P., Braman, N., Orooji, M., Rakshit, S., Bera, K., Rajiah, P., Ginsberg, J., et al.: Perinodular and intranodular radiomic features on lung ct images distinguish adenocarcinomas from granulomas. Radiology **290**(3), 783 (2019)
2. Bertram, R., et al.: Eye movements of radiologists reflect expertise in ct study interpretation: A potential tool to measure resident development. Radiology **281**(3), 805–815 (2016)
3. Bhattacharya, M., et al.: Radiotransformer: A cascaded global-focal transformer for visual attention-guided disease classification. arXiv preprint arXiv:2202.11781 (2022)
4. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. **35**, 99–109 (1943)
5. Bhattacharyya, A.: On a measure of divergence between two multinomial populations. Sankhyā: the indian journal of statistics pp. 401–406 (1946)
6. Carion, N., et al.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fogel, I., et al.: Gabor filters as texture discriminator. Biological cybernetics **61**(2), 103–113 (1989)
9. Van der Gijp, A., et al.: How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education **22**(3), 765–787 (2017)
10. Goldberger, A.L., et al.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), e215–e220 (2000)
11. Guo, Z., et al.: A completed modeling of local binary pattern operator for texture classification. IEEE transactions on image processing **19**(6), 1657–1663 (2010)

12. Hassani, A., et al.: Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704 (2021)
13. Jocher, G., et al.: yolov5. Code repository https://github. com/ultralytics/yolov5 (2020)
14. Johnson, A.E., et al.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
15. Karargyris, A., et al.: Eye gaze data for chest x-rays. PhysioNet https://doi.org/10.13026/QFDZ-ZR67 (2020)
16. Kelahan, L.C., et al.: The radiologist's gaze: mapping three-dimensional visual search in computed tomography of the abdomen and pelvis. Journal of digital imaging **32**(2), 234–240 (2019)
17. Kelly, B.S., et al.: The development of expertise in radiology: in chest radiograph interpretation,"expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification. Radiology **280**(1), 252–260 (2016)
18. Khosravan, N., et al.: A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning. Medical image analysis **51**, 101–115 (2019)
19. Lakhani, P., et al.: The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. (2021)
20. Lanfredi, R.B., et al.: Reflacx: Reports and eye-tracking data for localization of abnormalities in chest x-rays
21. Lanfredi, R.B., et al.: Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. arXiv preprint arXiv:2109.14187 (2021)
22. Lee, A., et al.: Identification of gaze pattern and blind spots by upper gastrointestinal endoscopy using an eye-tracking technique. Surgical Endoscopy pp. 1–8 (2021)
23. Li, Y., Shenoy, V., Prasanna, P., Ramakrishnan, I., Ling, H., Gupta, H.: Surgical phase recognition in laparoscopic cholecystectomy. arXiv preprint arXiv:2206.07198 (2022)
24. Lin, T.Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
25. Litchfield, D., et al.: Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. Journal of Experimental Psychology: Applied **16**(3), 251 (2010)
26. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
27. Nebbia, G., et al.: Radiomics-informed deep curriculum learning for breast cancer diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 634–643. Springer (2021)
28. Nguyen, H.Q., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. arXiv preprint arXiv:2012.15029 (2020)
29. Parekh, V.S., et al.: Deep learning and radiomics in precision medicine. Expert review of precision medicine and drug development **4**(2), 59–72 (2019)
30. Prasanna, P., et al.: Radiographic-deformation and textural heterogeneity (r-depth): an integrated descriptor for brain tumor prognosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 459–467. Springer (2017)
31. Redmon, J., et al.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

32. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
33. Singh, G., Manjila, S., Sakla, N., True, A., Wardeh, A.H., Beig, N., Vaysberg, A., Matthews, J., Prasanna, P., Spektor, V.: Radiomics and radiogenomics in gliomas: a contemporary update. British Journal of Cancer **125**(5), 641–657 (2021)
34. Tokuyasu, T., Iwashita, Y., Matsunobu, Y., Kamiyama, T., Ishikake, M., Sakaguchi, S., Ebe, K., Tada, K., Endo, Y., Etoh, T., et al.: Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. Surgical endoscopy **35**(4), 1651–1658 (2021)
35. Tourassi, G., et al.: Investigating the link between radiologists' gaze, diagnostic decision, and image content. Journal of the American Medical Informatics Association **20**(6), 1067–1075 (2013)
36. Venjakob, A., et al.: Radiologists' eye gaze when reading cranial ct images. In: Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment. vol. 8318, pp. 78–87. SPIE (2012)
37. Waite, S., et al.: Analysis of perceptual expertise in radiology–current knowledge and a new perspective. Frontiers in human neuroscience **13**, 213 (2019)
38. Wang, X., et al.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
39. Yoshie, T., et al.: The influence of experience on gazing patterns during endovascular treatment: Eye-tracking study. Journal of Neuroendovascular Therapy pp. oa–2021 (2021)
40. Zhou, X., et al.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
41. Zimmermann, J.M., et al.: Quantification of avoidable radiation exposure in interventional fluoroscopy with eye tracking technology. Investigative Radiology **55**(7), 457–462 (2020)