Deep Multiple Instance Learning with Spatial Attention for ROP Case Classification, Instance Selection and Abnormality Localization

Xirong Li*, Wencui Wan†, Yang Zhou‡, Jianchun Zhao‡, Qijie Wei‡, Junbo Rong†, Pengyi Zhou†
Limin Xu†, Lijuan Lang†, Yuying Liu†, Chengzhi Niu†, Dayong Ding‡, and Xuemin Jin†
*Key Lab of DEKE, Renmin University of China, Beijing 100872, China
†The First Affiliated Hospital of Zhengzhou University, Henan 450052, China
‡Vistel AI Lab, Visionary Intelligence Ltd, Beijing 100872, China
Email: xirong@ruc.edu.cn, wanwencui82@126.com

***Abstract*—This paper tackles automated screening of Retinopathy of Prematurity (ROP), one of the most common causes of visual loss in childhood. Clinically, ROP screening per case requires multiple color fundus image instances that capture different zones of the (premature) retina. A desirable model shall not only make a decision at the case level, but also pinpoint which instances and what part of the instances are responsible for the decision. This paper makes the first attempt to accomplish three tasks, *i.e.* ROP case classification, instance selection and abnormality localization in a unified framework. To that end, we propose a new model that effectively combines instance-attention based deep multiple instance learning (MIL) and spatial attention (SA). The propose model, which we term MIL-SA, identifies positive instances in light of their contributions to case-level decision. Meanwhile, abnormal regions in the identified instances are automatically localized by the SA mechanism. Moreover, MIL-SA is learned from case-level binary labels exclusively, and in an end-to-end manner. Experiments on a large clinical dataset of 2,186 cases with 11,053 fundus images show the viability of the proposed model for all the three tasks.**

***Index Terms*—ROP screening, deep multiple instance learning, spatial attention, instance selection, abnormality localization**

I. Introduction

Retinopathy of prematurity (ROP), caused by disorganized growth of retinal blood vessels [1], is a potentially blinding eye disease that affects preterm infants. As one of the most common causes of visual loss in childhood [2], early diagnosis of ROP is crucial. Unsurprisingly, we see an increasing amount of deep learning based efforts towards automated screening of ROP [3]–[10].

Different from disease recognition for the mature retina that normally requires a single color fundus image [11]–[14], ROP diagnosis clinically requires multiple color fundus images capturing different zones of the premature retina. According to the International Committee for the Classification of Retinopathy of Prematurity (ICROP) guideline [15]–[17], an ROP examination per eye, known as a *case*, shall cover typical zones include posterior pole, upper / lower nasal, upper / lower temporal, and peripheral, see Fig. 1. Based on the examination of the multiple image instances, a case-level decision of ROP *positive* or *negative* is made, while the positive outcome may further lead to fine-grained classification of the severity. Hence, a desirable model towards automated ROP screening shall not only make a decision at the case level, but also pinpoint which instances and what part of the instances are responsible for the decision.

Trained on instance-level annotations, deep convolutional neural networks (CNN) have shown promising performance for normal / ROP classification [8], [10], normal / pre-plus / plus classification [4], [7], and multi-stage classification [9], all at the instance level. An intrinsic drawback of the above methods is the need of a large amount of well-labeled instances, the collection of which is known to be expensive and time consuming. Moreover, how to properly upgrade instance-level predictions to the case level is nontrivial [3].

Few attempts exist for learning directly from case-level annotations [3], [5], [6]. Worrall *et al.* [3] simply treat the label of a training case as the label of all its instances, and accordingly train an instance-level CNN. While being simple, such a strategy is questionable, as not all instances in a positive case are positive, see Fig. 1. To cope with the issue, deep multiple instance learning (MIL) is investigated in Hu *et al.* [6] and Wang *et al.* [5], reporting performance better than [3]. As illustrated in Fig. 2, a deep MIL framework can be conceptually decomposed into three blocks including instance-level feature extraction, feature aggregation, and case-level classification. In [5], [6], convolutional blocks of ResNet-50 [18] are used to extract features for individual instances, followed by a Max Pooling operation to obtain features for a given case. Despite their encouraging results, we notice two deficiencies in the current MIL-based method for ROP classification. First, it lacks a mechanism to measure the importance of the individual instances. Second, it lacks a means to localize abnormal regions within specific instances. While [6] employs Guided-BP [19], a gradient-based visualization algorithm, to show which regions are activated, no quantitative evaluation is provided. In fact, recent evidence in the context of image recognition shows that gradient-based visualization is suboptimal [20].

Towards automated and interpretable ROP screening, this

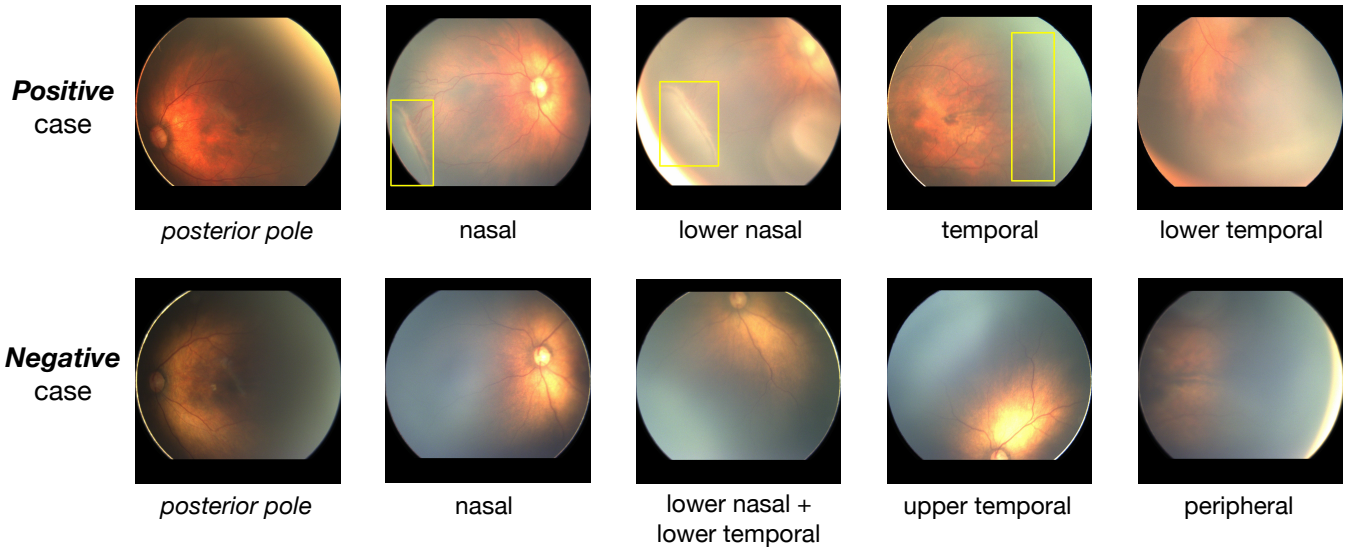

Fig. 1. **Positive / negative cases of Retinopathy of Prematurity (ROP)**. An ROP examination per eye requires color fundus imaging on distinct retinal quadrants, see text below each image. Yellow bounding boxes show abnormal regions (*extraretinal fibrovascular proliferation* here). Multiple image instances form a case, upon which this retinal disease is clinically diagnosed. A positive case may contain negative instances, see the first and last images of the first row. In this paper, given only training data with case-level binary labels, we not only aim for case classification, but also want to find out which instances and what part of the instances are responsible.

paper makes the following three contributions:

- To the best of our knowledge, this is the first work on tackling three tasks, *i.e.*, case classification, instance selection and abnormality localization in a unified framework.
- Inspired by the success of instance-attention based deep MIL [21], we propose MIL with Spatial Attention (MIL-SA), see Fig. 2. Weakly trained using only case-level binary labels, MIL-SA is capable of selecting instances deemed to be ROP positive and subsequently localizing abnormal regions in the selected instances.
- Experiments on a large clinical dataset of 2,186 cases with 11,053 color fundus images show the viability of the proposed MIL-SA model.

II. Proposed Approach

We aim to accomplish three subtasks related to automated ROP screening, namely case classification, instance selection and abnormality localization by a single model. Moreover, the model needs to be trained with case-level labels only. To that end, we propose a deep Multiple Instance Learning (MIL) approach. For the ease of consistent description, we denote a set of n color fundus images acquired by an ROP examination on a specific eye as $\mathcal{X} = \{x_1, \ldots, x_n\}$. For a specific image instance x of size $m \times m \times 3$ ($m = 256$ in this work), we use $x(i, j)$ to access its pixel at position $(i, j), i = 1, \ldots, m, j = 1, \ldots, m$. Let y be a binary variable indicating whether a case, an instance, or a pixel is ROP positive. A desirable model shall produce the following three posterior probabilities, *i.e.*

$$\begin{cases} P(y = 1|\mathcal{X}), \\ P(y = 1|x), \\ P(y = 1|x(i, j)), \end{cases} \tag{1}$$

which respectively indicate the probability of the case $\mathcal{X}$, the instance x, and the pixel $x(i, j)$ being ROP positive. Accordingly, case classification is performed by thresholding $P(y = 1|\mathcal{X})$ at 0.5. Instance selection is achieved by sorting $\{x_1, \ldots, x_n\}$ in descending order in terms of $P(y = 1|x)$. Last, abnormality localization is approached by unsupervised object localization on a $m \times m$ saliency map $\{P(y = 1|x(i, j), i = 1, \ldots, m, j = 1, \ldots, m)\}$.

A. MIL-SA: Deep Multiple Instance Learning with Spatial Attention

Our model is developed on the basis of the attention-based deep multiple instance learning network (MIL-att) by Ilse *et al.* [21]. So we first describe MIL-att in the new context, and then introduce our task-specific improvements.

As illustrated in Fig. 2, the deep MIL network conceptually consists of three blocks that work in a sequential manner. That is, 1) a CNN-based feature extraction block that extracts feature vectors per instance, 2) an instance-attention based feature aggregation block that compresses the n vectors into a single vector, and lastly 3) a case-level classification block. For a given case $\mathcal{X}$, its n instances are fed into the CNN, which is ResNet-101 [18] here, in a batch mode. Consequently, each instance x_i results in an array of feature maps of size $8 \times 8 \times 2048$, denoted as $F_i = \{f_{i,1}, \ldots, f_{i,2048}\}$. After Global Average Pooling (GAP), F_i is reduced to $\bar{F}_i$, a more compact vector of 1×2048. In order to aggregate $\{\bar{F}_1, \ldots, \bar{F}_n\}$ into

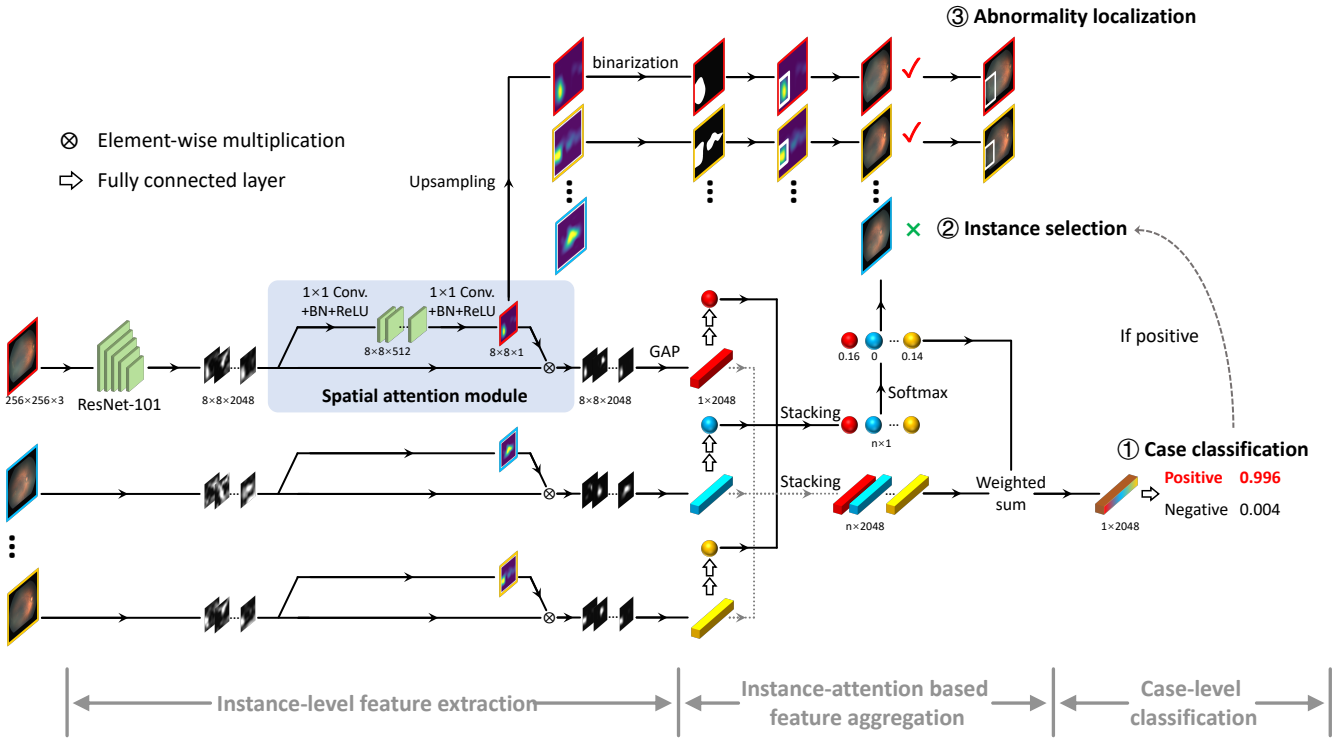

Fig. 2. **Conceptual diagram of the proposed MIL-SA model**, developed by combining an instance-attention based deep multiple instance learning network [21] and Spatial Attention (SA). While trained exclusively on case-level binary labels, MIL-SA is capable of accomplishing three tasks related to automated ROP screening, *i.e.*, case classification, instance selection and abnormality localization, in a unified framework. In particular, the spatial attention maps, produced by the SA module (Eq. 5), are used for re-weighing the feature maps from the ResNet-101 based backbone to produce instance-level attention weights, shown in colored circles. The SA maps are also used to localize abnormal regions by unsupervised object detection.

a case-level representation, Ilse *et al.* introduce an instance-based attention module. Instance-wise attention weights $\{a_i\}$ are obtained by the following self-attention network,

$$\{a_i\} = softmax(\{FC_{128\times1}(tanh(FC_{2048\times128}(\bar{F}_i)))\}), \tag{2}$$

where $i = 1, \ldots, n$ and $FC_{128\times1}$ indicates a fully connected layer which converts a 128-dimensional input vector to a 1-dimensional output vector. In a similar vein, we define $FC_{2048\times128}$. Hence, the self-attention network consists of two FC layers with $tanh$ as its intermediate activation function and $softmax$ to ensure that the attention weights $\{a_i\}$ sum to 1.

For obtaining a case-level feature representation, denoted by $\bar{F}_{att}$, instance-attention based feature aggregation is performed as a weighted sum of the instance features:

$$\bar{F}_{att} = \sum_{i=1}^{n} a_i \cdot \bar{F}_i, \tag{3}$$

The feature then goes to the classification block to predict whether the given case is ROP positive. That is,

$$\{P(y=1|\mathcal{X}), P(y=0|\mathcal{X})\} = softmax(FC_{2048\times2}(\bar{F}_{att})). \tag{4}$$

Since the attention weights measure the importance of the individual instances, we use them to implement $P(y=1|x)$, *i.e.*, $P(y=1|x_i) = a_i$.

Spatial attention module. In order to estimate the pixel-wise importance $P(y=1|x(i,j))$, we introduce spatial attention (SA) into the MIL framework, terming the new model MIL-SA. As shown in Fig. 2, our SA module consists of two 1×1 convolution layers, each followed by batch normalization and ReLU. Note that we use ReLU instead of the commonly used sigmoid activation [20], [22], as our preliminary experiment shows that sigmoid over-reduces the gap between active and inactive regions and thus results in less discriminative saliency maps. Given A_i as the spatial attention map produced by the SA module with respect to F_i and $F_{i,sa}$ as the updated feature maps, we have

$$\begin{aligned} A_i &= ReLU(BN(Conv_{1\times1}(ReLU(BN(Conv_{1\times1}(F_i)))))), \\ F_{i,sa} &= F_i \otimes A_i, \end{aligned} \tag{5}$$

where $\otimes$ indicates the Hadamard product. Given the new feature maps, $\bar{F}_i$ used in Eq. 2 and 3 is replaced by its SA-embedded counterpart, namely GAP($F_{i,sa}$). As a consequence, both the instance-based attention module and the SA module are jointly learned in an end-to-end manner to minimize the case classification loss. This, in turn, provides a measure of how each instance / pixel contributes to the final decision. Moreover, training the proposed MIL-SA model requires only case-level binary annotations to compute the cross-entropy loss based on Eq. 4.

In the inference stage, to localize the most abnormal region in a given instance, we first enlarge its spatial attention map A to $m \times m$ by bilinear upsampling. An unsupervised object localization algorithm [23] is conducted on the enlarged attention map. In particular, the map is binarized with a threshold of 15% of the max intensity, resulting in one or more connected segments of pixels. Abnormality localization is obtained by finding the minimal bounding box that encloses the largest segment.

III. Experiments

A. Experimental Setup

1) Data collection: Despite the increasing amount of research efforts on automated ROP screening [3], [4], [6], [8], [9], no dataset is publicly available. We therefore constructed an expert-labeled dataset as follows. The dataset was collected via 1,173 examinations for 333 infants, performed between 2014 and 2018 in a state eye center. The type of fundus cameras used was RetCam 3. An examination on a specific eye is treated as one case. We collected 2,186 cases with 11,053 image instances in total.

Expert annotation was collectively conducted by six ophthalmologists skilled in evaluating infant eyes. For each image, its regions corresponding to specific ROP lesions are marked out by bounding boxes. Five lesions specified by the ICROP guideline [17] are considered: *demarcation line* (DL), *ridge*, *extraretinal fibrovascular proliferation* (EFP), *partial retinal detachment* (PRD), and *vascular dilatation and tortuosity* (VDT). A fundus image is labeled as a *positive* instance if it shows a specific lesion complying with ICROP, and *negative* otherwise. Accordingly, a case is considered positive if it contains at least one positive instance, and negative otherwise. Discrepancies naturally exist between individual experts. So for a reliable evaluation, we pick 313 cases that receive consistent labels from three experts, and split at random into the validation and test sets. The remainder is used for training. Basic data statistics are summarized in Table I. Since PRD is rare, we exclude this lesion from our experiments. Note that all the instance and region annotations are merely used for evaluation.

2) Details of model training: All models evaluated in our experiments use ResNet-101 [18] pretrained on ImageNet as their backbones. We use SGD with momentum of 0.9, weight decay of 10^{-5}, batch size of 6, and initial learning rate of 10^{-3}. The learning rate is divided by 10 if the performance does not improve in 10 consecutive epochs. Once the rate reaches 10^{-7}, early stop occurs. Model selection is based on the AUC score of the validation set. Common data augmentation such as random flip, rotation, and changes in brightness and contrast is applied on instances. As for case-level data augmentation, we follow [6] to make each training case have a fixed number of eight instances.

B. Experiment 1. ROP Case Classification

1) Baselines: We compare with the prior art [6] which uses multiple instance learning with max pooling (MIL-max) for case classification. We also compare with Ilse *et al.* [21], where max pooling is replaced by an instance attention module (MIL-att). While MIL-att is shown to be more effective than MIL-max in several multiple instance learning tasks [21], its effectiveness has not been verified in the new context.

2) Evaluation criteria: We use Sensitivity, Specificity and the F1 score as their harmonic mean. We also report Accuracy (the ratio of correctly classified cases) and AUC.

3) Results: Performance of the three models is given in Table II. The proposed MIL-SA outperforms the two baselines in F1, accuracy and AUC. The result shows that adding spatial attention into the deep MIL network is helpful for ROP case classification.

C. Experiment 2. ROP Instance Selection

1) Baselines: We again compare MIL-SA with MIL-max and MIL-att. Notice that for both MIL-SA and MIL-att, attention weights $\{a_i\}$ associated with each instance in a given case can naturally be used as a ranking criterion for instance selection. In order to measure the importance of the individual instances for MIL-max, we use Grad-CAM [23], a state-of-the-art technique for visually interpreting decisions made by a specific CNN. Specifically, Grad-CAM generates a saliency map of 8×8 for each instance by convolving its feature maps F_i with the back-propagated gradient maps, followed by ReLU. The instance weight is obtained by a Global Average Pooling operation on the saliency map.

2) Evaluation criteria: As instance selection is essentially a ranking task, we calculate Average Precision (AP) [24] per case. Overall performance is measured by averaging AP scores over all test cases.

3) Result: As Table III shows, Both MIL-att and MIL-SA outperform MIL-max + Grad-CAM. Since AP is positively correlated to the positive rate, we divide the test cases into distinct groups based on their positive rates, reporting mean AP per group. The performance gap tends to be larger for groups with lower positive rates. These result indicate that the instance-attention mechanism estimates the instance importance more accurately than the ad-hoc combination of MIL-max and Grad-CAM. Again, MIL-SA is found to be better than the baselines.

D. Experiment 3. Abnormality Localization

1) Baselines: Unlike MIL-SA, MIL-max and MIL-att have no spatial attention module to measure the importance of specific regions in each instance. We re-use the saliency maps of MIL-max from the previous experiment. In a similar vein, we use Grad-CAM to generate saliency maps for MIL-att. The abnormal region per positive instance is localized on the saliency maps, using the unsupervised object localization algorithm described in Section II.

2) Evaluation criterion: We report Intersection over Union (IoU) between predicted boxes and ground-truth boxes, as commonly used for (weakly supervised) object detection [23], [25].

TABLE I

STATISTICS OF DATA USED IN OUR EXPERIMENTS FOR ROP CASE CLASSIFICATION, INSTANCE SELECTION AND ABNORMALITY LOCALIZATION. NOTE THAT OUR MIL-SA NETWORK IS LEARNED FULLY FROM CASE-LEVEL BINARY LABELS, WHILE ANNOTATIONS W.R.T INSTANCES AND LESIONS ARE USED ONLY FOR DATA ANALYTICS. AS (PARTIAL) *retinal detachment* RARELY OCCURS IN OUR DATASET, WE EXCLUDE THE LESION FROM THIS STUDY. DL, EFP, AND VDT STAND FOR *demarcation line*, *extraretinal fibrovascular proliferation*, AND *vascular dilatation and tortuosity*, RESPECTIVELY.

Data split	**No. of cases**		**No. of instances**		**No. of instances with specific lesions**			
	Positives	*Negatives*	*Positives*	*Negatives*	*DL*	*ridge*	*EFP*	*VDT*
training	859	1,014	2,599	6,995	748	980	458	1,086
validation	51	112	251	488	31	75	124	135
test	41	109	191	529	25	73	73	90

TABLE II

PERFORMANCES OF DIFFERENT MODELS FOR ROP CASE CLASSIFICATION. THE PROPOSED MIL-SA OUTPERFORMS THE BASELINES IN TERMS OF F1, ACCURACY AND AUC.

Model	**Sensitivity**	**Specificity**	**F1**	**Accuracy**	**AUC**
MIL-max [6]	**0.9512**	0.9083	0.9292	0.9200	0.9758
MIL-att [21]	**0.9512**	0.9358	0.9434	0.9400	0.9655
proposed *MIL-SA*	0.9268	**0.9725**	**0.9491**	**0.9600**	**0.9895**

TABLE III

PERFORMANCE OF DISTINCT MODELS FOR ROP INSTANCE SELECTION. THE POSITIVE RATE PER CASE IS THE RATE OF POSITIVE INSTANCES IN A SPECIFIC CASE. THE PROPOSED MIL-SA PERFORMS THE BEST, SHOWING ADVANTAGES ESPECIALLY FOR CASES WITH LOWER POSITIVE RATES.

	Overall	**Range of the positive rate per case**			
		$(0, 0.3)$	$[0.3, 0.5)$	$[0.5, 0.7)$	$[0.7, 1]$
Case number (percentage)	41 (100%)	1 (2.4%)	4 (9.8%)	8 (19.5%)	28 (68.3%)
Models:					
MIL-max + Grad-CAM	0.8991	0.5833	0.8833	0.8534	0.9257
MIL-att	0.9694	**1.0**	0.9375	0.9172	0.9877
proposed *MIL-SA*	**0.9811**	**1.0**	**1.0**	**0.9302**	**0.9922**

3) Results: As Table IV shows, MIL-SA outperforms MIL-att + Grad-CAM and MIL-max + Grad-CAM. Some localization examples are given in Fig. 3. Consider the last column for instance. MIL-max + Grad-CAM finds blot hemorrhage, which is however irrelevant w.r.t ROP. By contrast, MIL-SA finds VDT. Both quantitative and qualitative results show that MIL-SA better localizes abnormal regions related to ROP.

TABLE IV

PERFORMANCE OF DIFFERENT MODELS FOR ROP ABNORMALITY LOCALIZATION. LARGER IOU SCORES MEAN MORE PRECISE LOCALIZATION.

Model	**Overall**	**DL**	**ridge**	**EFP**	**VDT**
MIL-max + Grad-CAM	0.2594	0.2489	0.2942	0.3562	0.2258
MIL-att + Grad-CAM	0.2956	0.2853	0.3573	0.3879	0.2046
proposed *MIL-SA*	**0.3615**	**0.3814**	**0.4023**	**0.4621**	**0.2374**

IV. CONCLUSIONS AND REMARKS

In this paper, we tackle ROP case classification, instance selection and abnormality localization in a unified MIL framework. This is achieved by the proposed MIL-SA model that is developed by extending an instance-attention based deep MIL network with the spatial attention mechanism. MIL-SA is trained using case-level binary labels exclusively. Experiments on a clinical dataset of 2,186 cases and 11,053 images justify the effectiveness of MIL-SA against the prior art for all the three tasks.

On our test set, the new model achieves an AUC score of 0.9895 for ROP case classification, an AP score of 0.9811 for instance selection and IoU of 0.3615 for abnormality localization. Among the three tasks, much work remains to be done for abnormality localization. Given that acquiring pixel-level annotations is extremely expensive, an interesting future work is how to improve the model with a relatively small amount of pixel-level annotations, say by hybrid learning algorithms.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China (No. 61672523), Beijing Natural Science Foundation (No. 4202033), Beijing Natural Science Foundation Haidian Original Innovation Joint Fund (No. 19L2062), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (No. 2018PT32029), CAMS Initiative for Innovative Medicine (CAMS-I2M, 2018-I2M-AI-001), and the Pharmaceutical Collaborative Innovation Research Project of Beijing Science and Technology Commission (No. Z191100007719002). Corresponding author: Wencui Wan.

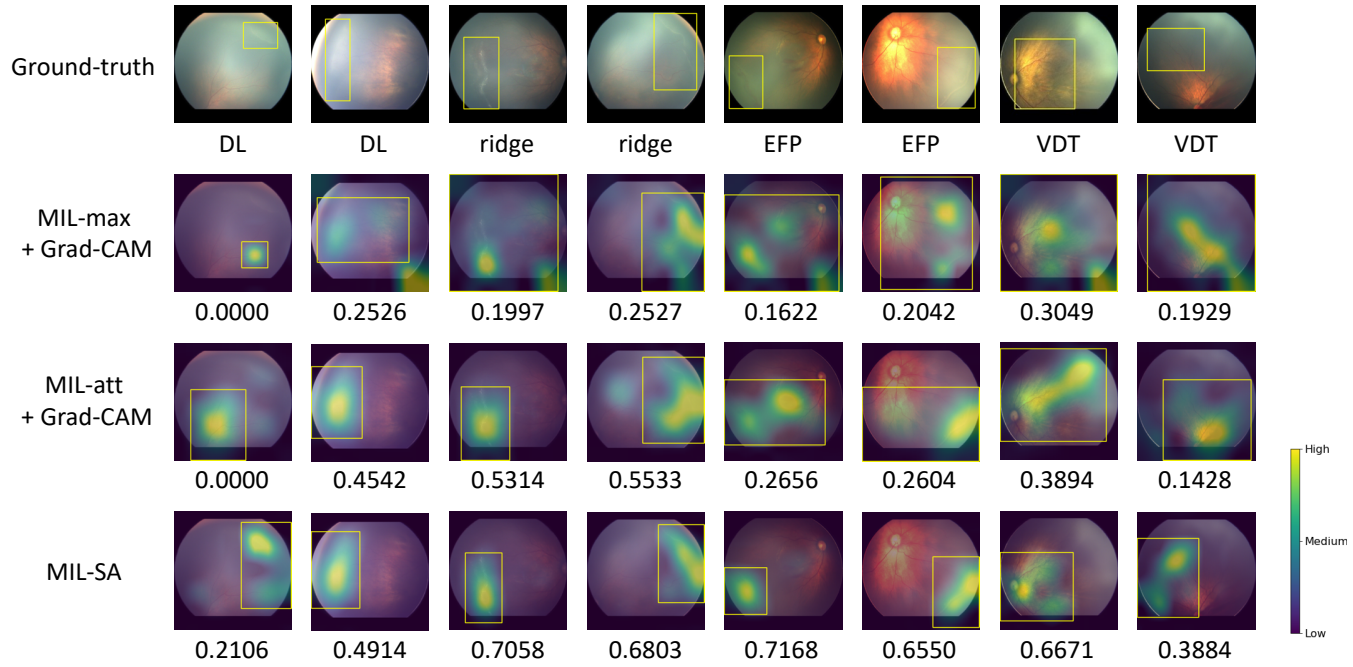

Fig. 3. **Examples of abnormality localization**. Numbers below saliency maps are IoU between regions localized by a specific model and ground truth. Larger is better. Compared to the two baselines, the proposed MIL-SA model localizes abnormal regions more accurately. Best viewed on screen.

REFERENCES

[1] P. Sapieha, J. S. Joyal, J. C. Rivera, E. Kermorvant-Duchemin, F. Sennlaub, P. Hardy, P. Lachapelle, and S. Chemtob, "Retinopathy of prematurity: Understanding ischemic retinal vasculopathies at an extreme of life," *Journal of Clinical Investigation*, vol. 120, no. 9, pp. 3022–3032, 2010.

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

[3] D. E. Worrall, C. M. Wilson, and G. J. Brostow, "Automated retinopathy of prematurity case detection with convolutional neural networks," in *Proceedings of LABELS/DLMIA@MICCAI*, 2016.

[4] J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. 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, pp. 803–810, 2018.

[5] J. Wang, R. Ju, Y. Chen, L. Zhang, J. Hu, Y. Wu, W. Dong, J. Zhong, and Z. Yi, "Automated retinopathy of prematurity screening using deep neural networks," *EBioMedicine*, vol. 35, pp. 361–368, 2018.

[6] J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 269–279, 2018.

[7] T. K. Redd, J. P. Campbell, J. M. Brown, S. J. Kim, S. Ostmo, R. V. P. Chan, *et al.*, "Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity," *British Journal of Ophthalmology*, vol. 103, no. 5, pp. 580–584, 2019.

[8] R. Zhang, J. Zhao, G. Chen, T. Wang, G. Zhang, and B. Lei, "Aggressive posterior retinopathy of prematurity automated diagnosis via a deep convolutional network," in *Proceedings of OMIA@MICCAI*, 2019.

[9] G. Chen, J. Zhao, R. Zhang, T. Wang, G. Zhang, and B. Lei, "Automated stage analysis of retinopathy of prematurity using joint segmentation and multi-instance learning," in *Proceedings of OMIA@MICCAI*, 2019.

[10] Y. Peng, W. Zhu, F. Chen, D. Xiang, and X. Chen, "Automated retinopathy of prematurity screening using deep neural network with attention mechanism," in *Proceedings of SPIE*, 2020.

[11] Q. Wei, X. Li, H. Wang, D. Ding, W. Yu, and Y. Chen, "Laser scar detection in fundus images using convolutional neural networks," in *Proceedings of ACCV*, 2018.

[12] X. Wang, L. Ju, X. Zhao, and Z. Ge, "Retinal abnormalities recognition using regional multitask learning," in *Proceedings of MICCAI*, 2019.

[13] W. Wang, Z. Xu, W. Yu, J. Zhao, J. Yang, F. He, Z. Yang, D. Chen, D. Ding, Y. Chen, and X. Li, "Two-stream CNN with loose pair training for multi-modal AMD categorization," in *Proceedings of MICCAI*, 2019.

[14] Q. Wei, X. Li, W. Yu, X. Zhang, Y. Zhang, B. Hu, B. Mo, D. Gong, N. Chen, D. Ding, and Y. Chen, "Learn to segment retinal lesions and beyond," in *Proceedings of ICPR*, 2020.

[15] International Committee for the Classification of Retinopathy of Prematurity, "An international classification of retinopathy of prematurity," *Archives of Ophthalmology*, vol. 102, no. 8, pp. 1130–1134, 1984.

[16] T. Aaberg *et al.*, "An international classification of retinopathy of prematurity: II. The classification of retinal detachment," *Archives of Ophthalmology*, vol. 105, no. 7, pp. 906–912, 1987.

[17] International Committee for the Classification of Retinopathy of Prematurity, "The international classification of retinopathy of prematurity revisited," *Archives of Ophthalmology*, vol. 123, no. 7, p. 991, 2005.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR*, 2016.

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

[20] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of CVPR*, 2019.

[21] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of ICML*, 2018.

[22] L. Mou, Y. Zhao, L. Chen, J. Cheng, Z. Gu, H. Hao, H. Qi, Y. Zheng, A. Frangi, and J. Liu, "CS-Net: Channel and spatial attention network for curvilinear structure segmentation," in *Proceedings of MICCAI*, 2019.

[23] 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 *Proceedings of ICCV*, 2017.

[24] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Computing Surveys*, vol. 49, no. 1, pp. 14:1–14:39, 2016.

[25] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L. B. Moreira, J. Eschbacher, P. Nakaji, M. C. Preul, and Y. Yang, "Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images," in *Proceedings of MICCAI*, 2018.