Research Article | Vol. 13, No. 8 / 1 Aug 2022 / Biomedical Optics Express 4087
Biomedical Optics EXPRESS

ADS-Net: attention-awareness and deep supervision based network for automatic detection of retinopathy of prematurity

YUANYUAN PENG,[1] ZHONGYUE CHEN,[1] WEIFANG ZHU,[1] FEI SHI,[1] MENG WANG,[2] YI ZHOU,[1] DAOMAN XIANG,[3] XINJIAN CHEN,[1,4,5] AND FENG CHEN[3,6]

[1]*MIPAV Lab, School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu Province, 215006, China*
[2]*Institute of High Performance Computing, A*STAR, Singapore*
[3]*Guangzhou Women and Children's Medical Center, Guangzhou, 510623, China*
[4]*State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, 215123, China*
[5]*xjchen@suda.edu.cn*
[6]*eyeguangzhou@126.com*

Abstract: Retinopathy of prematurity (ROP) is a proliferative vascular disease, which is one of the most dangerous and severe ocular complications in premature infants. Automatic ROP detection system can assist ophthalmologists in the diagnosis of ROP, which is safe, objective, and cost-effective. Unfortunately, due to the large local redundancy and the complex global dependencies in medical image processing, it is challenging to learn the discriminative representation from ROP-related fundus images. To bridge this gap, a novel attention-awareness and deep supervision based network (ADS-Net) is proposed to detect the existence of ROP (Normal or ROP) and 3-level ROP grading (Mild, Moderate, or Severe). First, to balance the problems of large local redundancy and complex global dependencies in images, we design a multi-semantic feature aggregation (MsFA) module based on self-attention mechanism to take full advantage of convolution and self-attention, generating attention-aware expressive features. Then, to solve the challenge of difficult training of deep model and further improve ROP detection performance, we propose an optimization strategy with deeply supervised loss. Finally, the proposed ADS-Net is evaluated on ROP screening and grading tasks with per-image and per-examination strategies, respectively. In terms of per-image classification pattern, the proposed ADS-Net achieves 0.9552 and 0.9037 for Kappa index in ROP screening and grading, respectively. Experimental results demonstrate that the proposed ADS-Net generally outperforms other state-of-the-art classification networks, showing the effectiveness of the proposed method.

1. Introduction

Retinopathy of Prematurity (ROP) is an ocular disease, which frequently occurs in premature babies with low birth weight (less than 1500 g) or born before 32 weeks of pregnancy and is considered to be one of the major causes of childhood blindness worldwide [1,2]. The shorter the gestation or the lighter the birth weight, the more likely that infant is to develop ROP. In addition, with the improvement of the survival rate of premature infants worldwide, ROP has become a problem that cannot be ignored in both developed and developing countries, especially in developing countries [3]. For example, there are about 2 million premature babies born annually in China, and it is conservatively estimated that about 20000 preterm infants suffer from ROP [4]. In addition, it is estimated that about 30000 premature infants annually are blind or suffer severe vision impairment due to ROP around the world [5].

#461411 https://doi.org/10.1364/BOE.461411
Journal © 2022 Received 14 Apr 2022; revised 17 Jun 2022; accepted 21 Jun 2022; published 5 Jul 2022

Biomedical Optics EXPRESS

At present, the prevention and treatment mode of ROP at home and abroad is that ophthalmologists obtain the fundus images of premature infants using the RetCam imaging system, which has been widely used due to its simple operation, wide-angle imaging, and high resolution [6]. Then, retinopathy of prematurity can be divided into five stages, plus disease, and aggressive posterior retinopathy of prematurity (AP-ROP), which is based on the international classification of ROP (ICROP) [7,8]. In addition, the ICROP also defines three ROP zones according to the symptom location in ROP, with each centered on the optic disc. Figure 1 shows six examples, including a normal fundus image and five fundus images of stage 1 to 5.

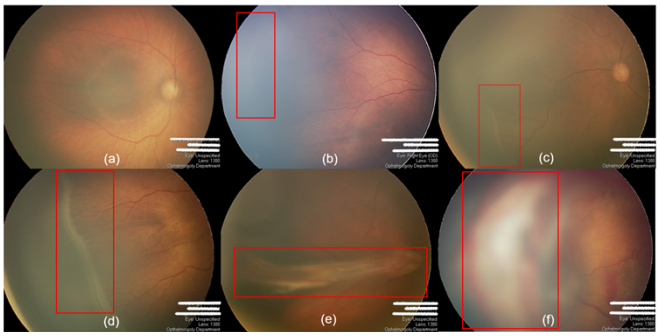

Fig. 1. Examples of normal and stage 1 to 5. The ROP lesion areas are in red boxes. (a) Normal. (b) stage 1. (c) stage 2. (d) stage 3. (e) stage 4. (f) stage 5.

Studies have shown that early screening, appropriate diagnosis, and timely treatment are the most effective ways to prevent blindness for premature infants with ROP [9]. Unfortunately, the global burden of the disease is still not fully addressed due to the following three main reasons. Firstly, barriers to screening include the lack of medical equipment and qualified professionals for ROP examination, especially in developing countries such as China and India [10,11]. Secondly, the images used for ROP diagnosis are usually blurry with low-contrast lighting, which may impair both human and machine interpretation of the images [12]. Eventually, accurate and objective diagnosis of ROP is difficult. Studies have shown that even among ROP clinical experts, the diagnostic variability is high due to their subjective interpretation, which may lead to significant differences in the clinical outcomes of preterm infants [13–15]. Based on the above factors, more and more researchers are interested in automatic ROP analysis and diagnosis based on artificial intelligence (AI) technology, which may improve the convenience and speed of ROP diagnosis and promote the standardization and objectivity of ROP diagnosis.

Over the few decades, inspired by the human multi-layer neural system, AI technology has made outstanding performance in medical image interpretation and diagnosis, such as the diagnosis of lung cancer [16], skin cancer [17] and breast histopathology [18] and the detection of glaucoma [19], age-related macular degeneration [20] and diabetic retinopathy [21]. In recent years, some studies have focused on ROP diagnosis, and have achieved promising results. Most of these studies focus on the identification of plus diseases, involving traditional methods and deep learning methods. The majority of traditional methods are to measure the statistics of the retinal blood vessels in the fundus images to identify plus disease, such as the diameter and curvature of the blood vessels. For example, "ROPTool" [22] and "i-ROP" [23] systems were developed to assist ophthalmologists in the diagnosis of plus disease, which required manually tracked and segmented vessels as input, thereby being limited in clinical applications. Zachary et al. developed a deep learning algorithm to automatically diagnose plus disease with high sensitivity and negative predictive value [24]. Driven by powerful deep neural networks and transfer learning in recent years, several deep learning based convolutional neural networks (CNNs) have been developed to focus on ROP screening, ROP severity grading, ROP staging,

Biomedical Optics EXPRESS

and ROP zoning. For example, pre-trained VGG16 was used to detect whether the ROP exists so as to realize the screening of ROP [25]. Hu et al. explored several architectures of CNNs pre-trained on ImageNet to detect the existence of ROP and grade the severity of ROP in a per-examination pattern [26]. Similar to the work of Hu et al., Huang et al. applied transfer learning to five deep neural network architectures to solve two problems in ROP detection: the existence of ROP (Normal or ROP) and the severity of ROP (mild-ROP or severe-ROP) [27]. Chen et al. used a fully convolutional network (FCN) to generate a binary segmentation map at the pixel level, which was then fed into a multi-instance learning (MIL) module along with the original image for 4-level ROP staging [28]. Ding et al. focused on diagnosing stages 1–3 by using object segmentation and convolutional neural network with transfer learning strategy [29]. Zhao et al. first used ResNet50 pre-trained on the Microsoft COCO dataset to detect the center of the optic disc and macula, then drew the boundary of zone I according to the ICROP standard, and finally judged the severity of the ROP disease [30]. To our best knowledge, this was the first time to use CNN to realize ROP zoning, but it can only identify zone I, not zone II and zone III. Recently, Agrawal et al. used U-Net and Hough circle transform to detect zones I, II and III, which involved optic disc and blood vessel segmentation [31]. In their method, macula's location was determined according to the Refs. [32] and [33] and repeated verification by senior ROP specialists. In addition, our previous works have focused on the ROP diagnosis, mainly solving three problems in ROP diagnosis. First, we used ResNet18 with transfer learning and attention mechanism for automatic ROP screening [34]. Second, we proposed a three-stream network with features fusion, transfer learning, and ordinal classification strategy for 5-level ROP staging in per-image and per-examination patterns [35]. Finally, a semi-supervised feature calibration adversarial learning network (SSFC-ALN) was proposed for 3-level ROP zoning [36]. In conclusion, CNNs based algorithms can realize automatic and objective ROP diagnosis, thus assist ophthalmologists in the diagnosis and treatment of ROP.

The current paper builds upon the previous successful models and proposes a novel methodology for the automatic ROP diagnosis. The method proposed in this paper is used to solve two ROP detection tasks. The first is to realize the screening of ROP, which is a binary classification problem. The second is to assess the severity of ROP according to 3-level ROP grading. Both tasks are analyzed in per-image and per-examination patterns. To realize the accurate ROP detection, several architectures based CNNs are explored, including the ResNext [37], DenseNet [38], ResNet [39], Inception [40], EfficientNet [41], HRNet [42], and ECA-Net [43], which have been proved to hold promise for ROP screening and grading in the following experiments of Section 3, especially for ROP screening. However, some challenges still exist on the ROP detection tasks. Firstly, the demarcation lines or ridges presented in the fundus images usually only account for a small part of the whole fundus image, which means that there is a lot of redundant information in the fundus image and this redundant information may affect the accuracy of ROP detection. Secondly, some ROP fundus images are characterized by the demarcation line between vascularized and avascular areas with significant differences in location and shape, which may lead to lower discriminability of the features learned by the hidden layer. Many previous studies [37–41] have shown that CNNs can effectively reduce local redundancy by convolution in a small neighborhood, but the limited receptive filed makes it difficult to capture global dependencies. Alternatively, self-attention mechanism can effectively capture long-range dependency. Therefore, to meet the first challenge, we design a novel attention module named as multi-semantic feature aggregation (MsFA) module based on self-attention mechanism, which can be embedded in CNNs to combine the advantages of convolution and self-attention. In addition, although increasing the depth of the network can improve the feature extraction capability of the network, it will also increase the difficulty of network optimization, which may lead to the disappearance of the gradient and the slow optimization speed. Previous studies have proved that deep supervised learning is helpful to solve the above problems [44–46], which has attracted our

Biomedical Optics EXPRESS

attention. Therefore, to meet the second challenge, the deep supervision strategy is adopted by adding auxiliary classifiers after some intermediate convolutional layers in our proposed network, which can not only accelerate convergence speed in deep neural network training, but also make full use of the feature information of hidden layers.

To sum up, we propose a novel attention-awareness and deep supervision based network (ADS-Net) for two ROP detection tasks in per-image and per-examination patterns. The main contributions of this paper can be summarized as follows:

(1) A novel attention module named MsFA module is designed and embedded into CNN, which can improve the ability of the model to capture the global long-range dependencies of multi-semantic features from different layers.

(2) Deep supervised learning is introduced by increasing auxiliary classifiers as supervision branches after some intermediate convolutional layers, which facilitates model training and allows the use of feature information of hidden layers for efficient model optimization.

(3) Both per-image and per-examination strategies are conducted to evaluate the proposed ADS-Net.

The remainder of this paper is organized as follows: the proposed method for automatic ROP screening and ROP grading is introduced in Section 2. Section 3 presents the experimental results in detail. In section 4, we conclude this paper and suggest future work.

2. Methodology

2.1. Overview framework

In this study, two ROP detection tasks are performed, including ROP screening and ROP grading. Figure 2 shows the proposed ADS-Net based ROP detection framework, consisting of a DenseNet121 for feature extraction, a MsFA module embedded into CNN to balance the problems of large local redundancy and complex global dependencies, and three classifiers to help with model optimization. Firstly, DenseNet121 is a dense convolutional network, making each layer in the network directly connected with its front layer to realize the reuse of features. In addition, DenseNet121 also has advantages in saving parameters and reducing overfitting [38]. Secondly, the MsFA module is embedded in the classification network to improve the ability of the model to capture the global long-range dependencies of multi-semantic features from different layers. Finally, to make full use of the feature information of hidden layers and help deep network optimization, the deep supervised learning strategy is adopted by adding two additional auxiliary classifiers after the second and third stages in DenseNet121 (classifier 1 and classifier 2 shown in Fig. 2). Notably, many previous studies have shown that transfer learning is an effective strategy to train the deep neural network when the target dataset is small [37–41], [47–49]. Therefore, transfer learning is used to help model training in this study.

2.2. Multi-semantic feature aggregation module

An important property of the human visual system is that one does not attempt to process the whole scene at once but selectively focuses on the salient parts to capture the visual structure better [50]. Inspired by this human perception process, many previous studies have widely explored attention mechanism, demonstrating that applying attention mechanism to convolutional neural network can increase its representation power to focus on important features, suppress the irrelevant ones and improve the performance in many computer vision tasks [43], [51–60]. For example, Hu et al. introduced a compact module named Squeeze-and-Excitation module (SE module) to exploit the inter-channel relationship, which applied global average-pooled features to computer channel-wise attention [52]. Recently, some new channel attention modules

Biomedical Optics EXPRESS

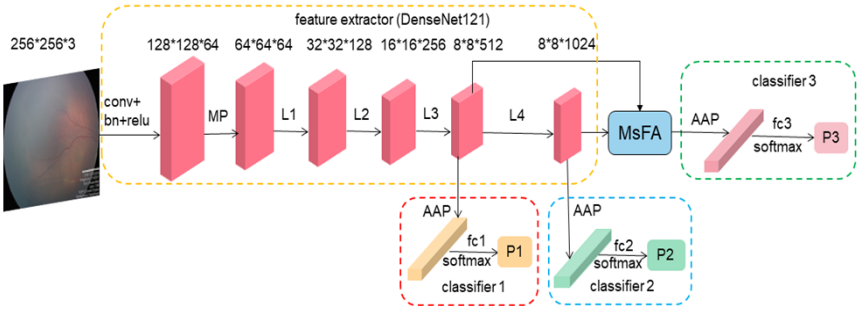

Fig. 2. An overview of the proposed ADS-Net based ROP detection framework. The ADS-Net consists of a feature extractor and three classifiers, where two auxiliary classifiers is in red and blue dotted boxes and a master classifier is in green dotted box. In addition, 'MP', 'AAP', 'fc' and 'softmax' represent max pooling operator, adaptive average pooling operator, fully connected operator, and Softmax activation layer, while 'L' represents multiple stacked dense connection module, 'P' represents the predicted classification results and 'MsFA' represents the proposed multi-semantic feature aggregation module as shown in Fig. 3.

have been proposed successively, such as Efficient Channel Attention module (ECA module), Pyramid Squeeze Attention module (PSA module), and Coordinate Attention module (Coord. Attention module) [43,54,55]. In addition, Woo et al. apply attention-based feature refinement with two distinctive modules from the channel and spatial axes and propose the Convolutional Block Attention Module (CBAM), which is a lightweight and general module [53]. Although such attention based feature extraction methods can improve the performance of CNNs, it still learns the feature relationships in limited receptive fields, which makes it hard to capture the global dependencies. Different from the above attention based methods, Fu et al. propose a Dual Attention Network (DANet) to capture rich context dependence based on the self-attention mechanism, which emphasizes meaningful features along the channel and spatial axes [51]. Therefore, considering the high redundancy and the complex global dependency of ROP fundus images, and previous studies have shown that CNNs and self-attention mechanism can alleviate the above problems respectively, we design a new attention module named as multi-semantic feature aggregation (MsFA) module as shown in Fig. 3 and embed it into the classification network based on the previous successful applications of CNNs and self-attention mechanism. As shown in Fig. 3, given two different input features $F_1 \in \mathbb{R}^{C',H',W'}$ and $F_2 \in \mathbb{R}^{C,H,W}$, where C and C' represent the channel numbers of two input features, H and H' represent height of two input features, and W and W' represent the width of two input features, the design of the MsFA module mainly consists of six steps:

(1) The convolution operation with 1×1 kernel size is adopted to transform the feature map F_1 into the same channel space as F_2, which is denoted as $F_1^{'} \in \mathbb{R}^{C,H',W'}$ $(H = H', W = W')$.

$$F_1' = \text{Conv1} \times 1(F_1) \in \mathbb{R}^{C,H',W'} \tag{1}$$

(2) A 1×1 convolutional operator is used to encode the input feature $F_1^{'} \in \mathbb{R}^{C,H,W}$ to generate a new feature map Q, and two 1×1 convolutional operators are used to encode the input feature $F_2 \in \mathbb{R}^{C,H,W}$ to generate two new feature maps K and V respectively, where $\{Q, K\} \in \mathbb{R}^{C/r,H,W}$ and $V \in \mathbb{R}^{C,H,W}$.

$$Q = \text{Conv1} \times 1(F_1') \in \mathbb{R}^{C/r,H,W} \tag{2}$$

$$K = \text{Conv1} \times 1(F_2) \in \mathbb{R}^{C/r,H,W} \tag{3}$$

$$V = \text{Conv1} \times 1(F_2) \in \mathbb{R}^{C,H,W} \tag{4}$$

Biomedical Optics EXPRESS

(3) We reshape and transpose Q to Q $\in \mathbb{R}^{H*W,C/r}$, and reshape K to K$\in \mathbb{R}^{C/r,H*W}$ and V to V $\in \mathbb{R}^{C,H*W}$, where C, H, and W represent the channel numbers, height and width of the input feature and r is the compression ratio.

$$Q = \text{Transpose}(\text{Reshape}(Q)) \in \mathbb{R}^{H*W,C/r} \tag{5}$$

$$K = \text{Reshape}(K) \in \mathbb{R}^{C/r,H*W} \tag{6}$$

$$V = \text{Reshape}(V) \in \mathbb{R}^{C,H*W} \tag{7}$$

(4) We do a matrix multiplication between Q and K and use a Softmax activation function to calculate the similarity matrix $E \in \mathbb{R}^{H*W,H*W}$ between query and key, as follows:

$$E = \text{Softmax}(Q*K) \in \mathbb{R}^{H*W,H*W} \tag{8}$$

where $*$ is the matrix multiplication operation.

(5) We do a matrix multiplication between V and the transpose of E to obtain the spatial response F_T and reshape it to $F_T \in \mathbb{R}^{C,H,W}$.

$$F_T = \text{Reshape}(V*E^T) \in \mathbb{R}^{C,H,W} \tag{9}$$

6) Finally, we do an element-wise summation between the original input feature F_2 and the above spatial response F_T to obtain the final spatial attention output $F_f \in \mathbb{R}^{C,H,W}$ as follows:

$$F_f = F_2 + \sigma * F_T \in \mathbb{R}^{C,H,W} \tag{10}$$

where σ is initialized to 0 and is a learnable parameter, which gradually learns to assign more weights. As can be seen from Eq. (10), the final feature map F_f is the weighted sum of the multi-semantic and strong semantic global features, which can adaptively build long-range dependencies from distant regions.

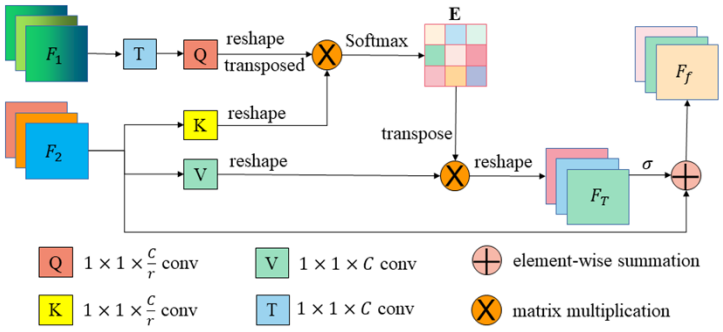

Fig. 3. Multi-semantic feature aggregation (MsFA) module. 'F_1' and 'F_2' represent two different input features of MsFA module, where 'F_1' and 'F_2' come from 'L_3' and 'L_4' in Fig. 2, respectively. 'E', 'F_T' and 'F_f' represent the similarity matrix, spatial response matrix and output feature, respectively. 'T' is a transform operation, which is obtained by a 1×1 convolution as shown in Eq. (1). In addition, 'Q', 'K' and 'V' are similar to the three branches of self-attention mechanism (query, key and value), which are realized by three 1×1 convolutions as shown in Eq. (2), (3) and (4).

Biomedical Optics EXPRESS

2.3. Deep supervised learning

In recent years, deep networks have brought performance gains in computer vision tasks. However, many empirical evidences suggest that the increase of network depth may increase the difficulty of network optimization and the performance improvement cannot be achieved by simply stacking more layers [61], which may cause the problem of gradient disappearance and slow convergence speed. Previous studies have shown that deep supervised learning can alleviate the above problems [44–46]. Therefore, to help network training and further improve the accuracy of ROP detection, the deep supervision strategy is adopted by increasing auxiliary classifier after some intermediate convolutional layers in our classification network. The auxiliary classifier as a branch of the network can provide extra supervision to the backbone network so as to enhance the representation ability of the network. In addition, if the network is too shallow, it cannot extract features with sufficient discriminating ability. Therefore, based on the above enlightenment, the auxiliary classifier branches are added in the later layers of the network to deepen the network with deep supervision strategy. As can be seen from Fig. 2, we add two auxiliary classifiers as supervision branches after the second and third stages in DenseNet121, so that the information of hidden layers can be used and different losses can be responsible for different parts of the weight layers in the network, which can propagate effective information through backward propagation. Similar to the previous study on deep supervision [44–46], three loss functions, including one master branch loss and two auxiliary losses are trained to pass through all previous layers of the backbone network. The two auxiliary losses are helpful for the learning process optimization, while the master branch loss takes the major responsibility. To achieve this, we add weights to balance the auxiliary losses and master loss as shown in Eq. (11). It is worth noting that in the test phase, we only use the optimized master branch to obtain the final ROP detection prediction.

2.4. Loss functions

In this study, we develop an attention-aware and deep supervision based network (ADS-Net) for ROP detection. The 3-channel fundus images with corresponding labels are fed into the ADS-Net to train the classifier. In addition, the strategy of deep supervision introduces the two auxiliary classification losses. Based on the analysis, the total loss function is defined as follow:

$$L = \alpha * L_1 + \beta * L_2 + \gamma * L_3 \tag{11}$$

$$L_j = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} I(y_i = k)\log(p(k|x_i)) \tag{12}$$

where L_j is the classification loss of j-th classifier ($j = 1,2,3$). α, β, and γ are three hyper-parameters and are set to 0.2, 0.3, and 0.5, respectively. m is the number of samples per mini-batch, y_i denotes the class label of image $x_i \cdot I(\cdot)$. is an indicator function, which equals one if y_i is equal to k, and zero otherwise.

3. Experiments and results

In this section, we first introduce the datasets and evaluation metrics for ROP detection. Then, we present the implementation details, including data preprocessing and the parameter settings in the training phase. Finally, we will give the detailed experimental results and the corresponding analysis.

3.1. Datasets and evaluation metrics

3.1.1. Datasets

In our experiments, we evaluate the proposed ADS-Net on two ROP detection datasets, which are both from Guangzhou Women and Children's Medical Center. The first one includes 7396

Biomedical Optics EXPRESS

fundus images from 1543 examinations, which is used to detect the existent of ROP (Normal or ROP). The other is used for 3-level ROP grading (Mild, Moderate, or Severe), which includes 1337 fundus images from 363 examinations. Both of them with a resolution of $640 \times 480 \times 3$ were collected using RetCam3 by professionals between 2012 to 2015. The collection and analysis of image data were approved by the Institutional Review Board of Guangzhou Women and Children's Medical Center and adhered to the tenets of the Declaration of Helsinki. An informed consent was obtained from the guardians of each subject to perform all the imaging procedures.

Data annotation was performed by three pediatric ophthalmologists, including one chief ophthalmologist with more than fifteen years of ROP clinical experience and two attending ophthalmologists with over three years of ROP clinical experience from Guangzhou Women and Children's Medical Center. The labeling of ROP is based on the symptoms described in the guide of ICROP. The process of annotation is divided into two phases. The ophthalmologists first annotated the images into normal or ROP types, followed by labeling 3-level ROP grading. All the annotation work was conducted by three pediatric ophthalmologists independently, and only the data with consistent results were used to evaluate the proposed network. According to the annotation result in the second phase, a high level of data imbalance is observed, where there are relatively few ROP data in stage 1 and 5. Therefore, to alleviate the data imbalance problem, we classify stage 1 and 2 as mild ROP, stage 3 as moderate ROP, and stage 4 and 5 as severe ROP. This type of ROP grading is based on previous studies [26,62,63], but the difference is that we subdivide stage 3 into moderate ROP because stage 3 is an important stage between demarcation line growth and retinal detachment. In addition, a patient usually includes one or more examinations, and after data annotation, one or more fundus images are usually included in one examination. The datasets used for training, validating, and testing the proposed ADS-Net are randomly split according to the examination of left and right eyes of each patient shown in Table 1 and Table 2. It is easy to observe from Table 1 and Table 2 that the number of the first dataset is more than that of the second dataset. There are two possible reasons. First, ophthalmologists have high consistency in the first annotation phase. In contrast, in the second annotation phase, high diagnostic variability is observed among them due to subjective evaluation, which has been proved in previous studies [14,64]. Second, for the ROP grading task, the fundus images of normal are not involved.

Table 1. The first dataset used for training, evaluating, and testing the proposed method in this study

Dataset	Normal		ROP		Total	
	examination	image	examination	image	examination	image
Training	595	2230	328	2211	923	4441
Validating	198	743	111	738	309	1481
Testing	197	739	114	735	311	1474
Total	990	3712	553	3684	1543	7396

3.1.2. Evaluation metrics

To quantitatively evaluate the performance of our ADS-Net on two ROP detection tasks, different metrics are calculated. For the first ROP detection task, which is a binary classification, five different metrics are calculated, including accuracy, recall, precision, F1-score, and Kappa [65,66]. The ROP grading task is a multi-classification problem, and the dataset categories are a bit unbalanced, as shown in Table 2. Therefore, same as previous studies [35,36], four common classification metrics, including weighted-average recall (W_R), weighted-average precision (W_P), weighted-average F1 score (W_F1), and Kappa, are introduced.

Biomedical Optics EXPRESS

Table 2. The second dataset used for training, evaluating, and testing the proposed method in this study

Dataset	Mild		Moderate		Severe		Total	
	examination	image	examination	image	examination	image	examination	image
Training	121	360	59	269	18	118	198	747
Validating	42	113	18	75	13	70	73	258
Testing	51	153	29	106	12	73	92	332
Total	214	626	106	450	43	261	363	1337

Table 3. Performance of different methods on image-based normal and ROP binary classification

Methods	Accuracy	Recall	Precision	F1-score	Kappa	Parameters (M)	FLOPs
DenseNet169 [38]	0.9735	0.9701	0.9767	0.9734	0.9471	12.4895	4.46e9
ResNet50 [39]	0.9728	0.9701	0.9754	0.9727	0.9457	23.5121	5.38e9
ResNext50 [37]	0.9728	0.9673	0.9780	0.9726	0.9457	22.9840	5.57e9
SE_ResNet50 [52]	0.9722	0.9333	0.9806	0.9719	0.9444	26.0431	5.08e9
SE_ResNext50 [52]	0.9701	0.9660	0.9739	0.9699	0.9403	25.5150	5.57e9
InceptionV4 [40]	0.9715	0.9701	0.9727	0.9714	0.9430	41.1459	8.34e9
EfficientB2 [41]	0.9654	0.9537	0.9763	0.9649	0.9308	7.7095	**2.56e7**
HRNet [42]	0.9769	0.9687	**0.9849**	0.9766	0.9539	19.2541	5.65e9
ECA-Net152 [43]	0.9735	0.9673	0.9793	0.9733	0.9471	58.1481	1.51e10
Peng et al. [34]	0.9729	0.9660	0.9793	0.9726	0.9457	17.6205	4.02e9
Zhang et al. [25]	0.9687	0.9619	0.9752	0.9685	0.9376	134.2637	2.02e10
Baseline	0.9664	0.9619	0.9658	0.9652	0.9308	**6.9559**	3.76e9
Baseline + MsFA	0.9735	0.9660	0.9807	0.9733	0.9471	8.6620	3.87e9
Baseline + DS	0.9708	**0.9728**	0.9688	0.9708	0.9417	6.9574	3.76e9
ADS-Net	**0.9776**	0.9714	0.9835	**0.9774**	**0.9552**	8.6636	3.87e9

3.2. Implementation and details

3.2.1. Data preprocessing

All fundus images are resized to $256 \times 256 \times 3$ by bilinear interpolation to reduce the computational cost. In addition, to eliminate the effects of different scales and illuminations, pixel intensity normalization is performed using the transforms.ToTensor module and transforms.Normalize module of a Pytorch graphics library called torchvision. In addition, we perform data augmentations for the two ROP detection tasks, including random rotation, horizontal and vertical flipping.

3.2.2. Parameter setting

The proposed ADS-Net is implemented based on the PyTorch framework. A NVIDIA GTX Titan X GPU with 12GB memory is used to train the model with back-propagation algorithm by minimizing the loss function as shown in Eq. (11). The optimizer is Adam, where both initial learning rate and weight decay are set to 0.0001 to optimize all networks. For the first ROP detection task, the batch size and epoch are set to 48 and 100, respectively, while for the ROP grading task, the batch size and epoch are set to 32 and 100. The compression ratio r in MsFA module is set to 16 in this study. During training, all networks are trained with identical optimization schemes, and we save the best model on the validation set. The code of the proposed ADS-Net will be released in: https://github.com/yuanyuanpeng0129/ADS-Net.

Biomedical Optics EXPRESS

3.3. Comparison experiments

In this paper, we propose an image-based ADS-Net for ROP automatic detection, including ROP screening and 3-level ROP grading. In addition, in the actual clinical diagnosis, the ophthalmologists usually browse multiple fundus images in one examination and take the diagnosis result of the most severe image as the final diagnosis result. Therefore, we also use an image-based classifier to verify the classification performance of our model according to examinations. Next, a series of comparative experiments and ablation experiments are presented and analyzed in detail.

3.3.1. Comparison experiments on ROP screening

For ROP screening, which is binary classification between normal and ROP, we evaluate the proposed ADS-Net on the testing dataset shown in Table 1, which contains 1474 fundus images from 311 examinations. Table 3 and Table 4 show the quantitative results of different methods in per-image and per-examination patterns, respectively. We compare the proposed ADS-Net with other excellent CNN based classification networks, including DenseNet169 [38], ResNet50 [39], InceptionV4 [40], EfficientNetB2 [41], ResNext50 [37], SE_ResNet50 [52], SE_ResNext50 [52], HRNet [42], and ECA-Net [43]. For convenience, we call the basic DenseNet121 pre-trained on ImageNet as the Baseline method. As can be seen from Table 3 and Table 4, our ADS-Net achieves superior performance in terms of most metrics.

Table 4. Performance of different methods on examination-based Normal and ROP binary classification

Methods	Accuracy	Recall	Precision	F1-score	Kappa
DenseNet169 [38]	0.9581	0.9737	0.9174	0.9447	0.9111
ResNet50 [39]	0.9453	0.9649	0.8943	0.9283	0.8842
ResNext50 [37]	0.9421	0.9561	0.8934	0.9237	0.8772
SE_ResNet50 [52]	0.9421	0.9474	0.9000	0.9231	0.8767
SE_ResNext50 [52]	0.9550	0.9649	0.9167	0.9402	0.9041
InceptionV4 [40]	0.9421	0.9561	0.8934	0.9237	0.8772
EfficientB2 [41]	0.9325	0.9561	0.8720	0.9121	0.8575
HRNet [42]	0.9518	0.9561	0.9169	0.9356	0.8971
ECA-Net152 [43]	0.9550	0.9737	0.9098	0.9407	0.9045
Peng et al. [34]	0.9518	0.9737	0.9024	0.9367	0.8978
Zhang et al. [25]	0.9486	0.9474	0.9153	0.9310	0.8900
Baseline	0.9357	0.9649	0.8730	0.9167	0.8645
Baseline + MsFA	0.9518	0.9649	0.9091	0.9362	0.8975
Baseline + DS	0.9518	**0.9737**	0.9024	0.9367	0.8978
ADS-Net	**0.9614**	0.9649	**0.9322**	**0.9483**	**0.9175**

As shown in Table 3, compared to the Baseline, our proposed ADS-Net improves the accuracy, recall, precision, F1-score, and Kappa by 1.16%, 0.99%, 1.83%, 1.26%, and 2.62%, respectively. Then, compared to other state-of-the-art classification networks, the performance of ADS-Net also gets more or less improvement in terms of most metrics with comparable or less model parameters and computational cost. For example, compared with the network with suboptimal classification performance (HRNet), the proposed ADS-Net with less model parameters and FLOPs (see the parameters and FLOPs in Table 3) improves the performance and achieves 0.9776 for accuracy, 0.9714 for recall, 0.9835 for precision, 0.9774 for F1-score, and 0.9552 for Kappa. While compared to EfficientNetB2 with comparable model parameters (see the parameters in

Biomedical Optics EXPRESS

Table 3), our method introduces more computational costs (see the FLOPs in Table 3), but also has a great improvement, which shows the effectiveness of our proposed ADS-Net. In addition, for the examination-based classification of Normal/ROP, a similar phenomenon can also be observed from Table 4. It is worth noting that the proposed ADS-Net is also compared with the two recent studies on the ROP screening [25,34], and the comparison experiments use the exact same ROP screening dataset in this study. The network proposed by Zhang et al. and our previous work are per-image classifiers, which are ImageNet pre-trained VGG16 and ImageNet pre-trained ResNet18 with attention mechanisms. As can be seen from Table 3, the proposed ADS-Net outperforms the other two methods on all metrics, which improves the Kappa by 1.88% and 1.00% compared with the other two methods, respectively. Moreover, the model parameters and computational cost of the proposed ADS-Net is less than that of the above two methods. The experimental results show the effectiveness of the proposed ADS-Net for the screening of ROP.

3.3.2. Comparison experiments on ROP grading

For ROP grading, we validate the proposed ADS-Net on 332 fundus images of ROP from 92 examinations. Table 5 and Table 6 give the quantitative results of different methods. As can be observed from Table 5 and Table 6, SE_ResNext50 [52] achieve suboptimal results, while ECA-Net152 [43] achieves the worst result in most metrics, especially in examination-based ROP grading. In addition, compared to the Baseline, the proposed ADS-Net with a slight increase in model parameters and computational cost gets an overall improvement for both image-based and examination-based ROP grading, which increases by 3.16% and 7.24% for the W_R of image-based and examination-based ROP grading. It can be seen from Table 5 and Table 6 that the proposed ADS-Net achieves the best performance in the image-based and examination-based ROP grading than other CNN-based classification methods. ADS-Net reaches 0.8886, 0.8880, 0.8883, and 0.9037 in terms of the W_R, W_P, W_F1, and Kappa for image-based ROP grading. In particular, compared with SE_ResNext50, which has the best performance among all the other methods for comparison, all indices of ADS-Net have been improved both in image-based and examination-based ROP grading and FLOPs shows the proposed ADS-Net is more efficient than SE_ResNext50. The experimental results prove the effectiveness of the proposed ADS-Net for ROP grading in fundus images.

Table 5. Image-based ROP grading results with different methods

Methods	W_R	W_P	W_F1	Kappa
DenseNet169 [38]	0.8524	0.8539	0.8530	0.8733
ResNet50 [39]	0.8343	0.8353	0.8345	0.8600
ResNext50 [37]	0.8343	0.8422	0.8359	0.8568
SE_ResNet50 [52]	0.8584	0.8603	0.8592	0.8779
SE_ResNext50 [52]	0.8795	0.8824	0.8804	0.9020
InceptionV4 [40]	0.8524	0.8646	0.8539	0.8719
EfficientB2 [41]	0.8464	0.8468	0.8449	0.8707
HRNet [42]	0.8735	0.8749	0.8741	0.8902
ECA-Net152 [43]	0.8223	0.8309	0.8246	0.8459
Baseline	0.8614	0.8743	0.8632	0.8774
Baseline + MsFA	0.8765	0.8800	0.8773	0.8922
Baseline + DS	0.8735	0.8846	0.8747	0.8882
ADS-Net	**0.8886**	**0.8880**	**0.8883**	**0.9037**

Biomedical Optics EXPRESS

Table 6. Examination-based ROP zoning results with different methods

Methods	W_R	W_P	W_F1	Kappa
DenseNet169 [38]	0.7500	0.7656	0.7546	0.7122
ResNet50 [39]	0.7283	0.7517	0.7339	0.6947
ResNext50 [37]	0.7174	0.7516	0.7242	0.6762
SE_ResNet50 [52]	0.7500	0.7722	0.7554	0.7112
SE_ResNext50 [52]	0.7609	0.7922	0.7666	0.7593
InceptionV4 [40]	0.6957	0.7752	0.7008	0.6557
EfficientB2 [41]	0.7609	0.7681	0.7636	0.7246
HRNet [42]	0.7500	0.7722	0.7554	0.7112
ECA-Net152 [43]	0.7065	0.7540	0.7134	0.6633
Baseline	0.7500	0.7881	0.7560	0.7092
Baseline + MsFA	0.7609	0.8032	0.7651	0.7251
Baseline + DS	0.7717	0.8093	0.7759	0.7366
ADS-Net	**0.8043**	**0.8215**	**0.8082**	**0.7671**

3.4. Ablation experiments

3.4.1. Ablation experiments for MsFA module

To prove the effect of the proposed MsFA module, we conduct a series of ablation experiments for ROP screening and ROP grading as shown in Table 3 and Table 5. Firstly, for ROP screening, the proposed MsFA module embedded in the Baseline (Baseline + MsFA) with a small increase in the number of model parameters and computational cost (see the parameters and FLOPs in Table 3) achieves improvement in terms of all evaluation metrics (0.73% for accuracy, 0.43% for recall, 1.54% for precision, 0.84% for F1-score and 1.75% for Kappa) as shown in Table 3. Secondly, for ROP grading, embedding the MsFA module into the Baseline (Baseline + MsFA) also obtains better classification performance. As shown in Table 5, compared with the Baseline, Baseline + MsFA gets an overall improvement, increasing by 1.75%, 1.57%, 1.63%, and 1.69% for the W_R, W_P, W_F1, and Kappa, respectively. Meanwhile, it can also be observed from Table 4 and Table 6 that the examination-based test results of the above two ROP detection tasks also show performance improvement. Taking the ROP grading for example, compared with Baseline, embedding the MsFA module into the Baseline (Baseline + MsFA) can improve the W_R, W_P, W_F1, and Kappa by 1.45%, 1.92%, 1.20%, and 2.24% respectively. The results demonstrate the effectiveness of the proposed MsFA module.

3.4.2. Ablation experiments for deep supervised learning

We adopt the deep supervised learning strategy by adding two auxiliary classifiers. Previous studies [44–46] have shown that the introduction of auxiliary losses can help optimize the learning process without affecting the learning of the master branch. To prove this point, we also conduct a series of ablation experiments. As can be seen from Table 3 and Table 5, compared with the Baseline, the introduction of deep supervised learning has improved the classification performance for two ROP detection tasks. Taking the image-based ROP grading for example, compared with the Baseline, the introduction of deep supervised learning strategy (Baseline + DS) improves the W_R, W_P, W_F1, and Kappa by 1.40%, 1.18%, 1.33%, and 1.23%, and achieves 0.8735 for W_R, 0.8846 for W_P, 0.8747 for W_F1, and 0.8882 for Kappa. A similar phenomenon can be observed in examination-based ROP grading from Table 6. There are two possible reasons for this performance improvement. Firstly, deep supervision can improve the directness and transparency of the hidden layers learning process, thereby improving the

Biomedical Optics EXPRESS

discrimination and robustness of features. Secondly, the two auxiliary losses play a regularization role, reducing the risk of overfitting.

4. Conclusion and discussions

In this paper, we propose a novel attention-awareness and deep supervision based network (ADS-Net) for ROP detection, including ROP screening and 3-level ROP grading. Firstly, to solve the challenge that ROP lesions account for a small proportion of the entire fundus image with complex global dependency as shown in Fig. 1, we design a novel attention module named as MsFA module, which is embedded into the later layers of the DenseNet121 to help model take full advantage of convolution and self-attention, promoting feature learning and producing more discriminative feature representation. Then, to solve the challenge of difficult training of deep model, the deep supervised learning strategy is adopted by adding two auxiliary classifiers after some intermediate convolutional layers in our ADS-Net. Finally, we conduct comprehensive experiments to validate the proposed ADS-Net in image-based and examination-based classification patterns. The experimental results show that compared with other state-of-the-art CNN-based classification networks, the classification performance of the proposed ADS-Net has been improved for two ROP detection tasks.

In addition, there is a major finding from Table 3 to Table 5, where the performance on ROP screening is much higher than that of 3-level ROP grading, indicating the latter task is much more difficult for CNNs. This finding is consistent with the clinical diagnosis that even for experienced ophthalmologists, the recognition of ROP grading is a tricky problem. In addition, the possible reason for such a challenge is that the amount of available data of the former task (Normal/ROP) is much larger than that of the latter task (3-level ROP grading). Therefore, in future works, we will collect more ROP grading data with high-quality annotation to further improve the performance of the proposed ADS-Net and validate the effectiveness of the proposed ADS-Net.

Funding. National Natural Science Foundation of China (61622114, U20A20170); National Key Research and Development Program of China (2018YFA0701700).

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. J. Chen and L. E. H. Smith, "Retinopathy of prematurity," Angiogenesis **10**(2), 133–140 (2007).
2. S. J. Kim, A. D. Port, R. Swan, J. P. Campbell, R. V. P. Chan, and M. F. Chiang, "Retinopathy of prematurity: a review of risk factors and their clinical significance," Surv. Ophthalmol. **63**(5), 618–637 (2018).
3. Y. Zhang and G. Zhang, "A Domain-Specific Terminology for Retinopathy of Prematurity and Its Applications in Clinical Settings," J. Healthcare Eng. **2018**(2018), 1–6 (2018).
4. Q. Li, Z. Wang, R. Wang, H. Tang, H. Chen, and Z. Feng, "A prospective study of the incidence of retinopathy of prematurity in China: evaluation of different screening criteria," J. Ophthalmology **2016**(2016), 1–8 (2016).
5. H. Blencowe, J. E. Lawn, T. Vazquez, A. Fielder, and C. Gilbert, "Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010," Pediatr Res **74**(S1), 35–49 (2013).
6. C. Wu, R. A. Petersen, and D. K. VanderVeen, "RetCam imaging for retinopathy of prematurity screening," J. Am. Assoc. Pediatric Ophthalmology and Strabismus **10**(2), 107–111 (2006).
7. T. Aaberg, "An international classification of retinopathy of prematurity: II. The classification of retinal detachment," Arch Ophthalmol **105**(7), 906–912 (1987).
8. International Committee for the Classification of Retinopathy of Prematurity, "The international classification of retinopathy of prematurity revisited," Ophthalmology **128**(10), e51–e68 (2021).
9. Y. Chen, J. Feng, C. Gilbert, H. Yin, J. Liang, and X. Li, "Time at treatment of severe retinopathy of prematurity in China: Recommendations for guidelines in more mature infants," PLoS ONE **10**(2), e0116669 (2015).
10. R. J. Vartanian, C. G. Besirli, J. D. Barks, C. A. Andrews, and D. C. Musch, "Trends in the screening and treatment of retinopathy of prematurity," Pediatrics **139**(1), 30–38 (2017).
11. J. P. Campbell, M. C. Ryan, E. Lore, P. Tian, S. Ostmo, K. Jonas, R. P. V. Chan, and M. F. Chiang, "Diagnostic discrepancies in retinopathy of prematurity classification," Ophthalmology **123**(8), 1795–1801 (2016).

Biomedical Optics EXPRESS

12. U. Sevik, C. Kose, T. Berber, and H. Erdol, "Identification of suitable fundus images using automated quality assessment methods," J. Biomed. Opt. **19**(4), 046006 (2014).
13. B. W. Fleck, C. Williams, and E. Juszczak, *et al.*, "An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials," Eye **32**(1), 74–80 (2018).
14. A. Gschließer, E. Stifter, T. Neumayer, E. Moser, and U. Schmidt-Erfurth, "Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity," Am. J. Ophthalmology **160**(3), 553–560.e3 (2015).
15. M. F. Chiang, L. Jiang, R. Gelman, Y. E. Du, and J. T. Flynn, "Interexpert agreement of plus disease diagnosis in retinopathy of prematurity," Arch Ophthalmol **125**(7), 875–880 (2007).
16. B. Van Ginneken, "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning," Radiol Phys Technol **10**(1), 23–32 (2017).
17. 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 **542**(7639), 115–118 (2017).
18. B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. Van Ginneken, N. Karssemeijer, and J. van der Laak, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," J. Med. Imag **4**(04), 1–044516 (2017).
19. S. K. Devalla, Z. Liang, T. H. Pham, C. Boote, N. G. Strouthidis, A. H. Thiery, and M. J. Girard, "Glaucoma management in the era of artificial intelligence," Br J Ophthalmol **104**(3), 301–311 (2020).
20. H. Liu, D. W. K. Wong, H. Fu, Y. Xu, and J. Liu, "DeepAMD: Detect early age-related macular degeneration by applying deep learning in a multiple instance learning framework," In *Asian Conference on Computer Vision*, (Springer: Cham, Switzerland, 2018), 625–640.
21. S. Wang, X. Wang, Y. Hu, Y. Shen, Z. Yang, M. Gan, and B. Lei, "Diabetic retinopathy diagnosis using multichannel generative adversarial network with semi-supervision," IEEE Trans. Automat. Sci. Eng. **18**(2), 574–585 (2021).
22. D. K. Wallace, Z. Zhao, and S. F. Freedman, "A pilot study using "ROPtool" to quantify plus disease in retinopathy of prematurity," J. Am. Assoc. Pediatric Ophthalmology and Strabismus **11**(4), 381–387 (2007).
23. A. C. Esra, B. C. Veronica, C. J. Peter, B. Alican, K. C. Jayashree, P. Samir, J. Karyn, R. V. P. Chan, and O. Susan, "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," Trans. Vis. Sci. Tech. **4**(6), 5–16 (2015).
24. Z. Tan, S. Simkin, C. Lai, and S. Dai, "Deep Learning Algorithm for Automated Diagnosis of Retinopathy of Prematurity Plus Disease," Trans. Vis. Sci. Tech. **8**(6), 23 (2019).
25. Y. Zhang, L. Wang, Z. Wu, J. Zeng, Y. Chen, R. Tain, J. Zhao, and G. Zhang, "Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images," IEEE Access **7**(1), 10232–10241 (2019).
26. J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," IEEE Trans. Med. Imaging **38**(1), 269–279 (2019).
27. Y. Huang, S. Vadloori, H. Chu, E. Kang, W. Wu, S. Kusaka, and Y. Fukushima, "Deep Learning Models for Automated Diagnosis of Retinopathy of Prematurity in Preterm Infants," Electronics **9**(9), 1444 (2020).
28. 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 *International Workshop on Ophthalmic Medical Image Analysis*. (Springer, 2019), 173–181.
29. A. Ding, Q. Chen, Y. Cao, and B. Liu, "Retinopathy of prematurity stage diagnosis using object segmentation and convolutional neural networks," *International Joint Conference on Neural Networks*. (2020), 1–6.
30. J. Zhao, B. Lei, Z. Wu, Y. Zhang, and G. Zhang, "A Deep Learning Framework for Identifying Zone I in RetCam Images," IEEE Access **7**(1), 103530–103537 (2019).
31. R. Agrawal, S. Kulkarni, R. Walambe, and K. Kotecha, "Assistive Framework for Automatic Detection of All the Zones in Retinopathy of Prematurity Using Deep Learning," J. Digit Imaging **34**(4), 932–947 (2021).
32. E. Alvarez, M. Wakakura, Z. Khan, and G. N. Dutton, "The disc-macula distance to disc diameter ratio: a new test for confirming optic nerve hypoplasia in young children," J. Pediatr Ophthalmol Strabismus **25**(3), 151–154 (1988).
33. D. D. Silva, K. D. Cocker, G. Lau, S. T. Clay, A. R. Fielder, and M. J. Moseley, "Optic disk size and optic disk-to-fovea distance in preterm and full-term infants," Invest. Ophthalmol. Vis. Sci. **47**(11), 4683–4686 (2006).
34. Y. Peng, W. Zhu, F. Chen, D. Xiang, and X. Chen, "Automated retinopathy of prematurity screening using deep neural network with attention mechanism," In *Medical Imaging 2020: Image Processing*. (2020), 1131321–1131327.
35. Y. Peng, W. Zhu, Z. Chen, M. Wang, L. Geng, K. Yu, Y. Zhou, T. Wang, D. Xiang, F. Chen, and X. Chen, "Automatic Staging for Retinopathy of Prematurity with Deep Feature Fusion and Ordinal Classification Strategy," IEEE Trans. Med. Imaging **40**(7), 1750–1762 (2021).
36. Y. Peng, Z. Chen, W. Zhu, F. Shi, M. Wang, Y. Zhou, D. Xiang, X. Chen, and F. Chen, "Automatic zoning for retinopathy of prematurity with semi-supervised feature calibration adversarial learning," Biomed. Opt. Express **13**(4), 1968–1984 (2022).
37. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2017), 1492–1500.
38. G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2017), 4700–4708.
39. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2016), 770–778.

Biomedical Optics EXPRESS

40. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," In *International conference on machine learning.* (2015), 1–11.
41. M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," arXiv, preprint arXiv:1905.11946 (2019).
42. J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, and W. Liu, "Deep high-resolution representation learning for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021).
43. Q. Wang, B. Wu, P. Zhu, P. Li, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 1–12.
44. L. Wang, C. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," arXiv preprint arXiv:1505.02496, 2015.
45. C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," in *Artificial Intelligence and Statistics* (AIS, 2015), pp. 562–570.
46. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proceedings of the IEEE conference on computer vision and pattern recognition.* (2017), 2881–2890.
47. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016).
48. S. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010).
49. C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," In *International conference on artificial neural networks.* (Springer, 2018), 270–279.
50. J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," arXiv preprint arXiv:1412.7755, (2014).
51. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2019), 3146–3154.
52. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2018), 7132–7141.
53. S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," In *Proceedings of the European Conference on Computer Vision.* (2018), 3–19.
54. Z. Hu, K. Zu, J. Lu, Y. Zou, and D. Meng, "Epsanet: An efficient pyramid split attention block on convolutional neural network," arXiv preprint arXiv:2105.14447, 2021.
55. Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design. arXiv 2021," arXiv preprint arXiv:2103.02907, 2021.
56. X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel Networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2020), 1–12.
57. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2018), 7794–7803.
58. F. Wang, M. Jiang, C. Qian, S. Yang, and X. Tang, "Residual attention network for image classification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE, 2017), 1–10.
59. T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (IEEE, 2014), 2–10.
60. B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," IEEE Trans. Multimedia **19**(6), 1245–1256 (2017).
61. L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," In *European conference on computer vision*, (Springer, Cham, 2016), 467–482.
62. Committee for the Classification of Retinopathy of Prematurity, "An international classification of retinopathy of prematurity," Arch. Ophthalmol. **102**(8), 1130–1134 (1984).
63. C. A. Ricard, C. E. L. Dammann, and O. Dammann, "Screening tool for early postnatal prediction of retinopathy of prematurity in preterm newborns (step-rop)," Neonatology **112**(2), 130–136 (2017).
64. V. Bolón-Canedo, E. Ataer-Cansizoglu, D. Erdogmus, J. Kalpathy-Cramer, O. Fontenla-Romero, A. Alonso-Betanzos, and M. F. Chiang, "Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach," Comput. Methods and Programs in Biomed. **122**(1), 1–15 (2015).
65. J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," arXiv preprint cmp-lg/9602004. (1996).
66. M. L. McHugh, "Interrater reliability: the kappa statistic," Biochem Med **22**(3), 276–282 (2012).