Automatic Staging for Retinopathy of Prematurity with Deep Feature Fusion and Ordinal Classification Strategy

Yuanyuan Peng, Weifang Zhu, Zhongyue Chen, Meng Wang, Le Geng, Kai Yu, Yi Zhou, Ting Wang, Daoman Xiang, Feng Chen, and Xinjian Chen, *Senior Member, IEEE*

***Abstract*—Retinopathy of prematurity (ROP) is a retinal disease which frequently occurs in premature babies with low birth weight and is considered as one of the major preventable causes of childhood blindness. Although automatic and semi-automatic diagnoses of ROP based on fundus image have been researched, most of the previous studies focused on plus disease detection and ROP screening. There are few studies focusing on ROP staging, which is important for the severity evaluation of the disease. To be consistent with clinical 5-level ROP staging, a novel and effective deep neural network based 5-level ROP staging network is proposed, which consists of multi-stream based parallel feature extractor, concatenation based deep feature fuser and clinical practice based ordinal classifier. First, the three-stream parallel framework including ResNet18, DenseNet121 and EfficientNetB2 is proposed as the feature extractor, which can extract rich and diverse high-level features. Second, the features from three streams are deeply fused by concatenation and convolution to generate a more effective and comprehensive feature. Finally, in the classification stage, an ordinal classification strategy is adopted, which can effectively improve the ROP staging performance. The proposed ROP staging network was evaluated with per-image and per-examination strategies. For per-image ROP staging, the proposed method was evaluated on 635 retinal fundus images from 196 examinations, including 303 Normal, 26 Stage 1, 127 Stage 2, 106 Stage 3, 61 Stage 4 and 12 Stage 5, which achieves 0.9055 for weighted recall, 0.9092 for weighted precision, 0.9043 for weighted F1 score, 0.9827 for accuracy with 1 (ACC1) and 0.9786 for Kappa, respectively. While for per-examination ROP staging, 1173 examinations with a 4-fold cross validation strategy were used to evaluate the effectiveness of the proposed method, which prove the validity and advantage of the proposed method.**

***Index Terms*—Retinopathy of Prematurity, Feature Fusion, Ordinal Classification, Automatic Staging, Fundus Images.**

I. INTRODUCTION

RETINOPATHY of prematurity (ROP) is caused by the abnormal development and proliferation of immature retinal vessels, which is a blinding eye disease accounting for about 19% of the causes of blindness in children worldwide and is often seen in premature infants with low gestational weeks (less than 32 weeks) and low birth weight (less than 1500g) [1-3]. According to 47 studies from 27 low- and middle-income countries, about half of low birth weight (LBW) infants are preterm rather than one-third of the pre-1990s hypothesis [4-5]. In China, the incidence of ROP in LBW infants is 26.0% [6].

With the increasing number of high-risk children in the world, ROP screening for high-risk children becomes particularly important [7]. According to the guidelines described by the international classification of ROP (ICROP), abnormal retinas of prematurity mainly includes three zones, five stages of ROP and a type of ancillary illness called plus disease based on the location, extent and severity of disease [8-10]. Five stages of ROP are used to characterize the severity of ROP according to the appearance of the retinal vessels at the avascular-vascular junction, which are shown in Fig. 1. A detailed description of the symptoms is given in Table 1. In addition, a type of ancillary illness called “plus” disease along with ROP is proposed, which can be found at any stage of ROP and is characterized by increased dilation and tortuosity in retinal vessels.

Standardized screening, graded diagnosis and treatment are effective ways to reduce the blindness rate of ROP. ROP screening tools mainly include binocular indirect ophthalmoscope and wide-angle digital fundus photography system in clinic. Wide-angle digital fundus photography is widely used due to its simple operation and high-resolution image [11]. The diagnosis of ROP requires the use of a wide-angle digital fundus imaging system with high image quality such as Retcam3 to examine the fundus of prematurity from

This study was supported in part by the National Key R&D Program of China (2018YFA0701700) and part by the National Nature Science Foundation of China (U20A20170, 61622114). Yuanyuan Peng and Weifang Zhu contributed equally to this work. Corresponding authors: Feng Chen, Xinjian Chen.

Yuanyuan Peng, Weifang Zhu, Zhongyue Chen, Meng Wang, Le Geng, Kai Yu, Yi Zhou and Ting Wang are with the School of Electronics and Information Engineering and Medical Image Processing, Analysis and Visualization Lab, Soochow University, Jiangsu 215006, China (Email: yypeng@stu.suda.edu.cn, wfzhu@suda.edu.cn).

Daoman Xiang and Feng Chen are with the Guangzhou Women and Children Medical Center, Guangzhou 510623, China (Email:eyeguangzhou@126.com)

Xinjian Chen is with the School of Electronics and Information Engineering and the State Key Laboratory of Radiation Medicine and Protection, Soochow University, Jiangsu 215006, China (Email: xjchen@suda.edu.cn).

different angles. The images are then interpreted by experienced ophthalmologists to determine whether the ROP and/or plus disease are present.

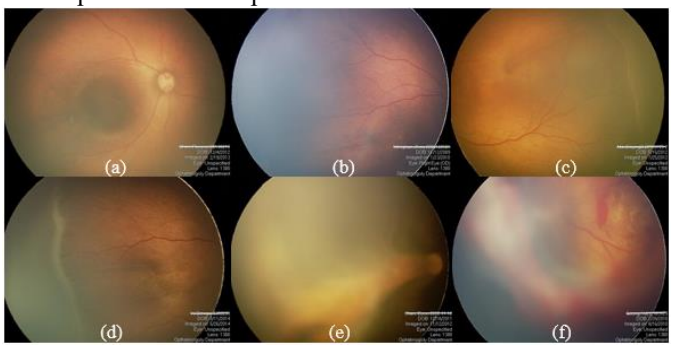


Fig. 1. Examples of normal and different stages of ROP. (a) Normal. (b) Stage 1. (c) Stage 2. (d) Stage 3. (e) Stage 4. (f) Stage 5.

TABLE I
SYMPTOMS OF STAGE 1 TO 5 OF ROP

Stage	Symptoms
1	A thin demarcation line that separates avascular retinal anteriorly from the vascular retinal posteriorly.
2	Line in stage 1 becomes wide and evolves to a ridge.
3	Extraretinal fibrovascular proliferation or neovascularization extends from the ridge into the vitreous.
4	Partial retinal detachment.
5	Total retinal detachment.

In the sensitive period of ROP, ophthalmologists carry out the routine fundus examination, early diagnosis and early treatment, and the success rate can be as high as 90%, which can effectively reduce the blindness rate. However, there are many difficulties in ROP screening, especially in developing countries [12-13]. First, the objective factor is the lack of medical equipment for ROP screening. Second, due to the complexity of professional knowledge for ROP, few ophthalmologists are qualified for ROP diagnosis. Furthermore, due to the subjective factors, ophthalmologists may be inconsistent in the ROP diagnosis, especially when plus disease is present [14-16]. Therefore, it is very important to develop a fast, objective and effective automatic ROP staging method.

To reduce the workload of ophthalmologists and improve the efficiency in ROP diagnosis, many computer-aided diagnosis systems have been proposed [17-24]. Most of the related work on automated or semi-automated methods for ROP diagnosis were focused on plus disease. For example, a system called "ROPTool" was proposed in [17] to assist ophthalmologists in diagnosing plus disease by quantitatively calculating tortuosity of vessels. E. Ataer-Cansizoglu et al. exploited principal spanning forest algorithm to develop a system named "i-ROP" [18], which was designed to grade plus disease into three types: normal, pre-plus and plus. In recent years, with the development of deep learnings, several studies have used ImageNet pre-trained DNNs for the ROP screening. For example, Hu et al. used Inception-V2 pre-trained on ImageNet combining with maximum aggregation operation to recognize the existence and severity of ROP from different fundus images in one examination [19-20]. Zhang et al. used VGG16 pre-trained on ImageNet for automated screening of ROP [21]. Our previous work used deep learning network with attention mechanism for automatic ROP screening [22]. Lei et al. utilized two deep convolution networks to automatically recognize aggressive posterior retinopathy of prematurity (AP-ROP), which is a retinal pathology characterized by sever vasodilation and distortion of the posterior pole of retina [23]. In addition, Chen et al. used joint segmentation and multi-instance learning for automatic ROP stage analysis including normal, stage 1, stage 2, stage 3 and stage 4 [24], which did not include stage 5.

However, the automatic 5-level ROP staging (stage1, stage2, stage3, stage4, and stage 5) has not been reported, which is important to the evaluation of the severity of the disease [1]. The main challenge of ROP staging is that the labeled fundus images are scarce and imbalanced, which is a common problem in medical image analysis, such as diabetic retinopathy diagnosis [25-26] and age-related macular degeneration analysis [27-28]. Many previous studies have demonstrated that feature-level fusion strategy can obtain much higher classification accuracy than general classification method. Inspired by the feature fusion for high-resolution aerial scene classification [29-37], the feature-level fusion strategy is utilized to further improve the performance in ROP staging in this study. In previous studies [19, 20, 21, 22,24], ROP staging was regarded as a standard multi-classification problem, in which the categories are assumed to be independent of each other. However, there is a strong ordinal relationship between categories in ROP staging, which is a gradual process from mild to severe. So considering that the cost of misclassification in clinical practice is different and inspired by [38-44], we utilize the ordinal classification for ROP staging, which can produce unequal punishment for different classification errors through loss function.

To sum up, we propose a novel three-stream deep network with feature-level fusion and ordinal classification strategy for the automatic 5-level ROP staging per-image and per-examination. The main contributions can be summarized as follows:

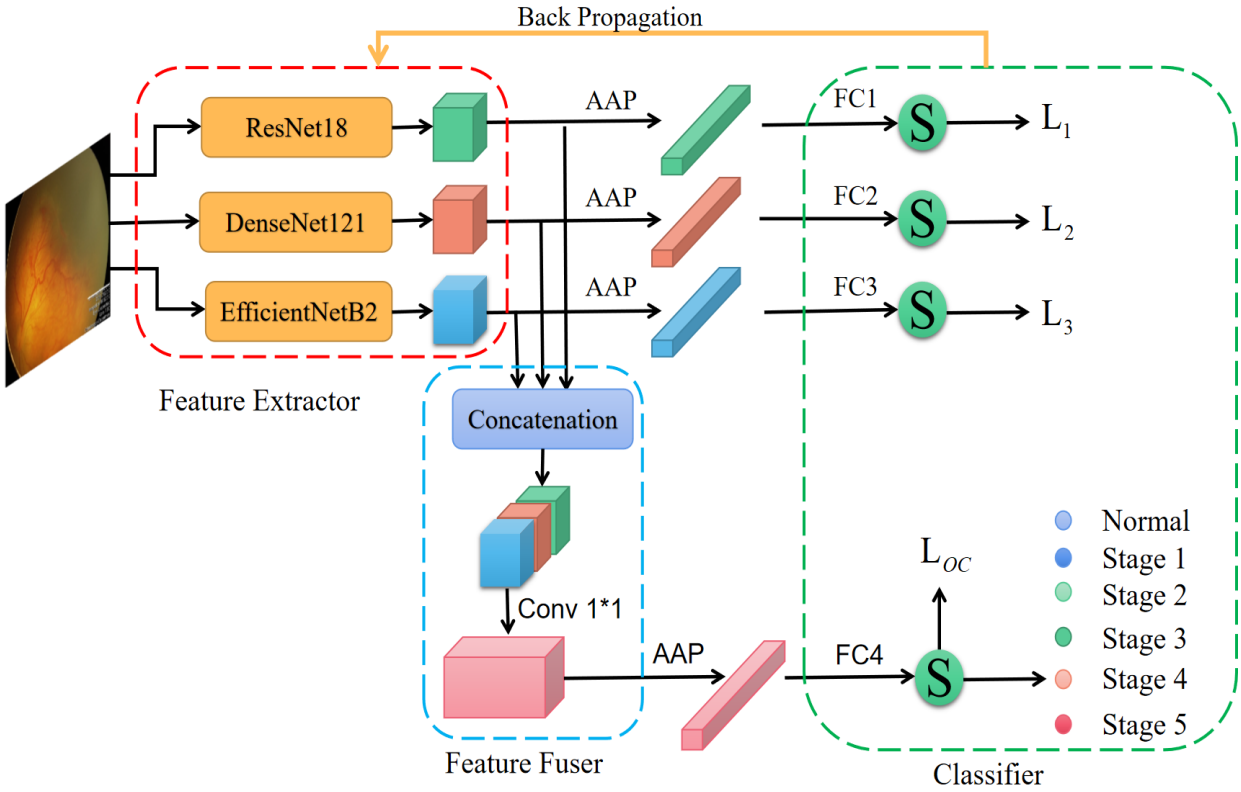
(1) A simple and effective framework consisting of three different parallel feature extraction deep networks is proposed for 5-level ROP staging.

(2) The concatenation method is used to fuse three high-level features extracted by three different parallel deep networks to obtain a richer and more effective feature for final staging.

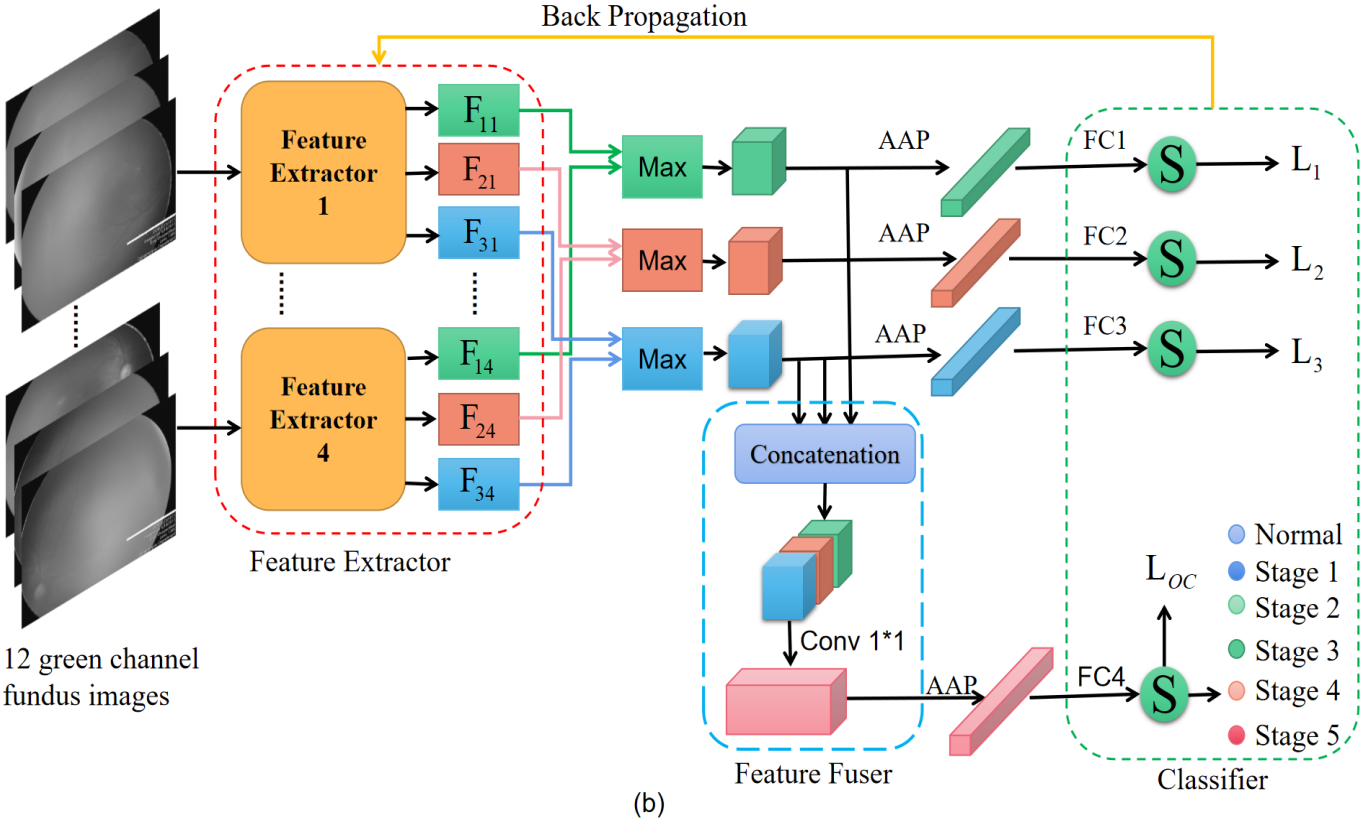
(3) The introduction of ordinal classification strategy into the convolution neural network improves the ROP staging performance.

(4) Both per-image and per-examination strategies are adopted in the evaluation of the proposed ROP staging network, which prove the effectiveness of our method.

The remainder of this paper is organized as follows: The proposed method for automatic ROP staging is introduced in Section II. Section III presents the experimental results in detail. In section IV, we conclude this paper and suggest future work.



(a)



(b)

Fig. 2. Schematic diagram of our proposed ROP staging networks. (a) per-image ROP staging network. (b) per-examination ROP staging network. Feature extractor, feature fuser and classifier are in the red, blue and green dotted boxes respectively. Feature extractor in red dotted boxes consists of ResNet18, DenseNet121 and EfficientNetB2. In the diagram, 'AAP', 'FC', 'S', 'L' and 'Max' represent adaptive average pooling operation, fully connected layer, sigmoid function, loss function and max feature aggressive operator, respectively.

II. METHODOLOGY

In this study, we perform ROP staging per-image and per-examination, respectively. For the former, the input of our network is an annotated fundus image. For the latter, the input of our network is an annotated examination containing multiple fundus images. Our proposed three-stream framework based ROP staging network for per-image is shown in Fig. 2 (a), which consists of ResNet18 [45], DenseNet121 [46] and EfficientNetB2 [47] for features extraction and feature-level fusion strategy for automatic ROP staging. The proposed network for per-examination ROP staging is shown in Fig. 2 (b), which is based on per-image ROP staging network shown in Fig. 2 (a). ResNet18 is a residual network with 18 weight layers, including convolution layer and full connection layer. It skillfully uses shortcut connection to transfer part of the original input feature information directly to the output feature, which simplifies the difficulty of feature learning, protects the integrity of feature information to a certain extent and solves the problem of model degradation in deep network [45]. DenseNet121 is a convolutional neural network with dense connections, which makes full use of shallow information, strengthens feature propagation, encourages feature reuse, reduces the number of network parameters greatly and can alleviate the problem of gradient disappearance [46]. EfficientNetB2 is an amplification network based on the baseline network EfficientNetB0 multiplied by constant rate. EfficientNetB0 was developed by leveraging a multi-objective neural architecture search that optimized both accuracy and FLOPS, whose main building block is mobile inversion bottleneck MBConvBlock [47]. It scales the model in the three dimensions including network depth, width and image resolution to balance the richness, fineness and information loss of the extracted features. Considering the order of each category in ROP staging, the idea of ordinal classification is adopted by modifying the label and the loss function of the multi-classification in the training. In the training stage, we use the transfer learning to obtain prior knowledge from the ImageNet dataset, which can accelerate network training and optimize network model [48-52]. In this section, the proposed ROP staging network will be illustrated in detail, including the network architecture, the feature-level fusion strategy, the ordinal classification and loss functions.

A. Three-Stream Feature Extraction Frameworks

In many previous studies about feature fusion [29-37], two-stream deep fusion framework was proposed, in which the two deep networks are the same. However, this may lead to the lack of diversity of the extracted features in our task. Inspired by [29-37], we propose a feature extractor as shown in Fig. 2, which contains three parallel different feature extractors including ResNet18, DenseNet121 and EfficientNetB2 to extract rich and diverse high level features expected to obtain richer and more effective features through the following feature fusion. There are two main reasons to develop such a feature extractor. First, different types of networks concern different types of features and can guarantee the diversity of the extracted feature, which is crucial to the classification performance. Second, because of the limited amount of ROP data, too complex models are easy to be overfitting, so these three relatively lightweight networks are selected for our ROP staging task. Theoretically，these three deep networks can be replaced by other ones according to the specific classification tasks.

As shown in Fig. 2 (b), the input for per-examination ROP staging is an examination containing multiple fundus images (12 images per-examination are adopted in this paper), which requires the network to predict the ROP stage according to the multiple fundus images comprehensively. First, 4 three-stream networks are used to extract features from 12 fundus images in the same examination in parallel. Then, inspired by researches [19-20], the max feature aggregated operator is adopted to obtain the maximum value of the 4 features from the same feature extractor.

TABLE II
THE LABELS ENCODING FORMS OF STANDARD CLASSIFICATION AND ORDINAL CLASSIFICATION

Attribution	Category	Standard Classification	Ordinal Classification
Normal	0	[1,0,0,0,0,0]	[1,0,0,0,0,0]
Stage 1	1	[0,1,0,0,0,0]	[1,1,0,0,0,0]
Stage 2	2	[0,0,1,0,0,0]	[1,1,1,0,0,0]
Stage 3	3	[0,0,0,1,0,0]	[1,1,1,1,0,0]
Stage 4	4	[0,0,0,0,1,0]	[1,1,1,1,1,0]
Stage 5	5	[0,0,0,0,0,1]	[1,1,1,1,1,1]

B. Feature Fusion

Many previous studies [29-37] have demonstrated that feature-level fusion of deep features for classification is a robust and effective strategy, which can combine N features extracted by N networks into a single feature vector that contains more discriminant information of the image. Previous research has used two common methods for feature fusion: parallel feature-level fusion strategy and serial feature-level fusion strategy. For the former, the fusion strategy requires features with the same dimensionality, and the common fusion methods include addition, maximum and mean operations, which are defined in Eq. (1), (2) and (3). For the latter, the dimensionalities of the features can be arbitrary and the features are fused by concatenation operation, which is shown in Eq. (4). The dimension of the fused feature vector is the sum of N features. In this study, the serial feature-level fusion strategy of concatenation is adopted.

$$F_{fusion} = \sum_{i=1}^{N} F_i \quad (1)$$

$$F_{fusion} = \max(F_1, F_2, \dots, F_N) \quad (2)$$

$$F_{fusion} = \frac{1}{N}\sum_{i=1}^{N} F_i \quad (3)$$

$$F_{fusion} = Concat(F_1, F_2, \dots, F_n) \quad (4)$$

where F_i is the i-th feature, N is the number of features, F_{fusion} is the fused feature and $Concat$ $(\cdot)$ denotes concatenation operation.

C. Ordinal Classification

The progression of ROP from mild to severe is a gradual process, whose staging characteristics are depicted in TABLE I. Just like the age estimation in [40] and ordinal sentiment analysis in [41], we consider the order of the categories in ROP staging and introduce ordinal classification.

Suppose D is an ordinal classification dataset with N samples (x_i,y_i) $(i=1,2,\ldots,N)$, where x_i is an input image and y_i is its corresponding ordinal category. Ordinal classification is different from standard classification due to the order of categories. For a standard multi-classification problem without considering the order of categories, the goal is to predict the probability of input image x_i belonging to category k ($y_i=k$). The category k encoded by one-hot labels is a vector $\mathbf{g} = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, where only g_k is set to 1 and all others are 0. Usually, softmax function is used to produce the probability distribution of outputs that the input image x_i belongs to each class. In the ordinal classification, the encoding of label is similar to the multi-label classification in form, but the labels are forced to be sorted. We list the label encoding forms of standard classification and ordinal classification in Table II. From the perspective of probability, it is to learn a mapping from the input image x_i to the output probability vector $P = \{p_1, p_2, \ldots, p_k, p_{k+1}, \ldots, p_t\}$, in which the target value of output nodes p_i $(i \le k)$ and p_i $(i > k)$ are 1 and 0, respectively. Thus, sigmoid function is usually used to produce the probability distribution of outputs. Although using independent sigmoid function for output nodes does not guarantee the monotonic relation $(p_1 \ge p_2 \ge, \ldots, \ge p_k \ge, \ldots, \ge p_t)$ [40], it is desirable for making predictions in our task.

D. Loss Function

Our proposed method is a parallel three-stream network consisting of three different deep networks which take the original fundus images as input for feature extraction and can obtain three different high-level features. Then, these three features are fused into a richer and more effective feature by concatenation, addition, mean or maximum operation. In particular, considering the order of categories and the data imbalance problem, the total loss function combined is defined as follow:

$$L = \alpha * L_{oc} + \sum_{i=1}^{3} \gamma_i * L_i \quad (5)$$

where,

$$L_{oc} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\beta_k * [I(t_i = k)\log\big(p(k|x_i)\big) + (1 - I(t_i = k))\log(1 - p(k|x_i))] \quad (6)$$

$$L_i = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\big[I(t_i = k)\log\big(p(k|x_i)\big) + \big(1 - I(t_i = k)\big)\log\big(1 - p(k|x_i)\big)\big], i = 1,2,3 \quad (7)$$

where,

$$\beta_k = \frac{single(k)}{total} \quad (8)$$

L_{oc} is the ordinal classification loss (K categories) in supervised learning where its corresponding label is as those listed in the fourth column of Table II. L_i is the standard classification loss (K categories) in supervised learning, in which its corresponding label is the form of one-hot listed in third column of Table II. γ_i (i=1,2,3) and α are four super-parameters and are all set to 0.25 in our experiments. m is the number of samples in per mini-batch, t_i denotes the class label of input image x_i. $I(\cdot)$ is an indicator function which equals one if t_i is equal to k (k =1, 2, …,K). β_k is the balance coefficient of class k in training process, $single(k)$ and $total$ are the numbers of class k and total number of training images.

III. Experiments And Results

In this section, the experimental dataset will be first described in detail. Then, we introduce the experimental setup, including the image processing and parameter settings in the training phase. Finally, the experimental results are presented in detail. A series of ablation studies are conducted to demonstrate the effectiveness of transfer learning, ordinal classification, feature fusion, and the proposed three-stream framework in ROP staging.

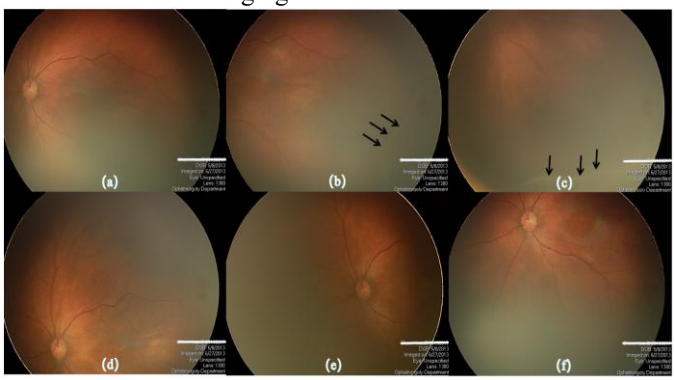


Fig. 3. Multiple fundus images from different shooting angles in an examination of a left eye. As indicated by the black marked arrows, an obvious ridge can be observed in (b) and (c) t, while (a), (d), (e) and (f) appear normally.

A. Data Imaging and Labeling

In this study, 9794 fundus images of 650 premature infants from 2024 ROP examinations were acquired using RetCam3 from the Guangzhou Women and Children Medical Center from 2012 to 2015. The collection and analysis of image data were approved by the Institutional Review Board of the Guangzhou Women and Children 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. The resolution of the images is 640 × 480 pixels. The number of images per-examination varies from 1 to 32, and the most frequent number is 6. The 6 images from an examination of a left eye in ROP stage 2 are shown in Fig. 3, in which the ridge in the fundus can be observed in Fig. 3 (b) and (c). The gestation age varies from 26 to 41 weeks, with a mean value of 32 weeks. 50% infants' gestation age is under 32 weeks and 42% of the infants' birth weight is less than 1500 grams.

The ground truth annotation is according to the symptoms described in Table I. 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 the Guangzhou Women and Children Medical Center participated in data labeling. Finally, 6110 fundus images with consistent labels among the three annotators are

included in our study, in which 4602 are normal fundus images and 1508 images from 508 examinations are abnormal ones ranging from stage 1 to 5. To balance the categories, 1639 normal fundus images from 665 examinations are randomly selected. So the final dataset used in our study contains 3147 fundus images from 1173 examinations. For per-image classification, all fundus images are divided into training set, validation set and testing set according to examinations, which are shown in the Table III. For per-examination classification, the numbers of examinations of each category are shown in the last row of Table III and a four-fold cross validation strategy is adopted to evaluate the ROP staging performance.

TABLE III
DATASET USED FOR TRAINING AND TESTING THE PROPOSED METHOD IN THIS STUDY

Dataset	Normal	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Total
Training	1034	73	287	269	229	60	1952
Validation	302	20	93	75	58	12	560
Test	303	26	127	106	61	12	635
Total images	1639	119	507	450	348	84	3147
Total examinations	665	64	181	146	89	28	1173

TABLE IV
PER-IMAGE CLASSIFICATION RESULTS WITH DIFFERENT METHODS

method	W_R	W_P	W_F1	ACC1	Kappa
ResNet18_1_Scratch [45]	0.5449	0.4775	0.4872	0.5906	0.2593
ResNet18_1 [45]	0.8362	0.8063	0.8202	0.9449	0.9424
DenseNet121_1 [46]	0.8488	0.8500	0.8453	0.9575	0.9530
EfficientNetB2_1 [47]	0.8409	0.8360	0.8332	0.9165	0.9283
ResNet18_1_OC	0.8614	0.8585	0.8572	0.9480	0.9610
DenseNet12_1_OC	0.8614	0.8674	0.8591	0.9591	0.9621
EfficientNetB2_1_OC	0.8441	0.8415	0.8395	0.9339	0.9409
ResNet18_3_Concatenation	0.8567	0.8671	0.8524	0.9591	0.9567
DenseNet121_3_Concatenation	0.8756	0.8857	0.8637	0.9717	0.9729
EfficientNetB2_3_Concatenation	0.8614	0.8583	0.8586	0.9307	0.9404
ResNet18_2_Concatenation	0.8535	0.8583	0.8442	0.9638	0.9468
ResNet18_4_Concatenation	0.8331	0.8104	0.8175	0.9764	0.9505
ResNet18_DenseNet121_Concatenation	0.8646	0.8679	0.8576	0.9543	0.9409
ResNet18_EfficientNetB2_Concatenation	0.8583	0.8286	0.8414	0.9638	0.9414
DenseNet121_EfficiebtNetB2_Concatenation	0.8598	0.8348	0.8441	0.9586	0.9470
TSF_Concatenation	0.8866	0.8966	0.8867	0.9732	0.9763
The proposed method	**0.9055**	**0.9092**	**0.9043**	**0.9827**	**0.9786**

'ResNet18_1_Scratch' represent the ResNet18 trained from scratch. 'ResNet18_1', 'DenseNet121_1' and 'EfficientNetB2_1' represent the single ResNet18, DenseNet121 and EfficientNetB2 pre-trained on ImageNet, respectively. 'ResNet18_1_OC', 'DenseNet121_1_OC' and 'EfficientNetB2_1_OC' represent the single ResNet18, DenseNet121 and EfficientNetB2 pre-trained on ImageNet with ordinal classification, respectively. 'ResNet18_2_Concatenation', 'ResNet18_3_Concatenation' and 'ResNet18_4_Concatenation' represent the two, three and four identical parallel ResNet18 pre-trained on ImageNet with concatenation feature fusion, respectively. 'DenseNet121_3_Concatenation' and 'EfficientNetB2_3_Concatenation' represent the three identical parallel ResNet18 and EfficientNetB2 pre-trained on ImageNet with concatenation feature fusion, respectively. 'ResNet18_DenseNet121_2_Concatenation', 'ResNet18_EfficientNetB2_2_Concatenation' and 'DenseNet121_EfficientNetB2_2_Concatenation' represent the pairwise combination of three different networks, respectively. 'TSF_Concatenation' and 'The proposed method' represent the three-stream framework using concatenation feature fusion without and with considering the order of categories, respectively.

TABLE V
PER-EXAMINATION CLASSIFICATION RESULTS WITH DIFFERENT METHODS ON PER-EXAMINATION CLASSIFICATION (MEAN ± STANDARD DEVIATION)

method	W_R	W_P	W_F1	ACC1	Kappa
ResNet18_1 [45]	0.7474±0.0272	0.6943±0.0170	0.7159±0.0202	0.8465±0.0105	0.8266±0.0283
DEnseNet121_1 [46]	0.7724±0.0184	0.7070±0.0189	0.7341±0.0204	0.8678±0.0097	0.8624±0.0114
EfficientNetB2_1 [47]	0.7543±0.0271	0.7604±0.0277	0.7533±0.0145	0.8611±0.0176	0.8813±0.0163
ResNet18_1_OC	0.7496±0.0280	0.7601±0.0387	0.7473±0.0310	0.8251±0.0128	0.8635±0.0132
DenseNet121_1_OC	0.7843±0.0151	0.7881±0.0101	0.7818±0.0153	0.8594±0.0207	0.8942±0.0164
EfficientNetB2_1_OC	0.7647±0.0307	0.7730±0.0277	0.7644±0.0305	0.8619±0.0213	0.8883±0.0096
TSF_Concatenation	0.7858±0.0202	0.7198±0.0203	0.7483±0.0204	0.8585±0.0225	0.8683±0.0248
The proposed method	**0.8006±0.0190**	**0.8039±0.0314**	**0.7958±0.0251**	**0.8706±0.0257**	**0.9107±0.0267**

B. Experimental Setup

1) Image Processing

To reduce the computational cost, all fundus images are down sampled to 256 ×256 using bilinear interpolation and normalized to [0,1]. For per-image classification pattern, to prevent over-fitting and enhance the generalization ability of the model, online data augmentation has been performed, including random rotation, horizontal flipping and vertical flipping. Previous studies [53-57] show that green channel of the fundus image has the highest contrast between retinal vessels and background. Therefore, considering the training speed, only green channel is used in per-examination classification. Consistent with previous studies [19,20], the image number of each examination is set to 12 in our study. For those examinations with less than 12 images, the images are randomly resampled to obtain 12, while for those with more than 12 images, 12 images are randomly selected as one examination.

As can be seen from Table III, the category distribution is unbalanced, where most ROP data are in stage 2 and 3, while stage 1, 4 and 5 are relatively few. There are two possible reasons: (1) Since both the boundary of stage 1 and the ridge of stage 2 are qualitative descriptions, there is subjectivity in the labeling process and ophthalmologists tend to label it as stage 2. (2) The phenomenon of partial and complete detachment of retina in stage 4 and 5 is rarely found in clinical practice, because effective treatment will be carried out before the disease progresses to stage 4 and 5 in most of the cases. To avoid the training problems caused by the imbalance of categories, balance coefficient β_k listed in Eq. (8) is introduced to modify the final loss function.

TABLE VI
PERFORMANCE COMPARISON ON PER-IMAGE TASK

Methods	W_R	W_P	W_F1	ACC1	Kappa
Inception-V4 [61]	0.8331	0.8443	0.8336	0.9480	0.9480
Inception-V4_OC	0.8457	0.8610	0.8449	0.9496	0.9528
ResNext50 [62]	0.8645	0.8749	0.8651	0.9543	0.9536
ResNext50_OC	0.8787	0.8802	0.8769	0.9601	0.9646
SE_ResNext50 [63]	0.8488	0.8547	0.8847	0.9559	0.9481
SE_ResNext50_OC	0.8835	0.8905	0.8847	0.9701	0.9630
SE_ResNet50 [63]	0.8567	0.8672	0.8551	0.9685	0.9523
SE_ResNet50_OC	0.8787	0.8843	0.8792	0.9528	0.9579
TST_Concatenation	0.8866	0.8966	0.8867	0.9732	0.9763
The proposed method	**0.9055**	**0.9092**	**0.9043**	**0.9827**	**0.9786**

TABLE VII
COMPARISON OF THE PROPOSED METHOD WITH HU ET AL.' METHOD WITH PER-EXAMINATION CLASSIFIER

Metrics	Hu et al. [19]	Proposed
W_R	0.7741±0.0244	0.8006±0.0190
W_P	0.7152±0.0340	0.8039±0.0314
W_F1	0.7305±0.0386	0.7958±0.0251
ACC1	0.8440±0.0174	0.8706±0.0457
Kappa	0.8523±0.0315	0.9107±0.0367

TABLE VIII
COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS WITH PER-IMAGE CLASSIFIER

Metrics	Zhang et al. [21]	Peng et al. [22]	Lei et al. [23]	Proposed
W_R	0.8724	0.8414	0.8519	0.9055
W_P	0.8797	0.8438	0.8095	0.9092
W_F1	0.8670	0.8387	0.9043	0.9043
ACC1	0.9685	0.9669	0.9827	0.9827
Kappa	0.9503	0.9532	0.9786	0.9786

2) Parameter Setting

The proposed three-stream framework is based on the ResNet18, DenseNet121, EfficientB2 pre-trained on ImageNet. The implementation is based on the PyTorch platform. We use a NVIDIA Tesla K40 GPU with 12GB memory to train the model with back-propagation algorithm by minimizing the loss function as shown in Eq. (5). Adam is used as the optimizer to minimize the loss function. Both initial learning rate and weight decay are set to 0.0001 to optimize the network. For per-image classification pattern, the batch size and epoch are set to 32 and 40, respectively. For per-examination classification pattern, the batch size and epoch are set to 8 and 60. During training, all networks are trained with identical optimization schemes and we save the best model on validation set.

3) Evaluation Metrics

Considering the category imbalance of the dataset shown in Table III, weighted recall (W_R), weighted precision (W_P), weighted F1 score (W_F1), accuracy within 1 (ACC1) [39] and Kappa index [20, 58-59] are introduced to evaluate the ROP staging performance. ACC1 is similar to accuracy, which allows a wider range of outputs as "right". For example, if the ground truth is stage 4, then the predictions of both stage 4 and

5 will be considered as right in ACC1, which is consistent with the common clinic principle of ROP staging.

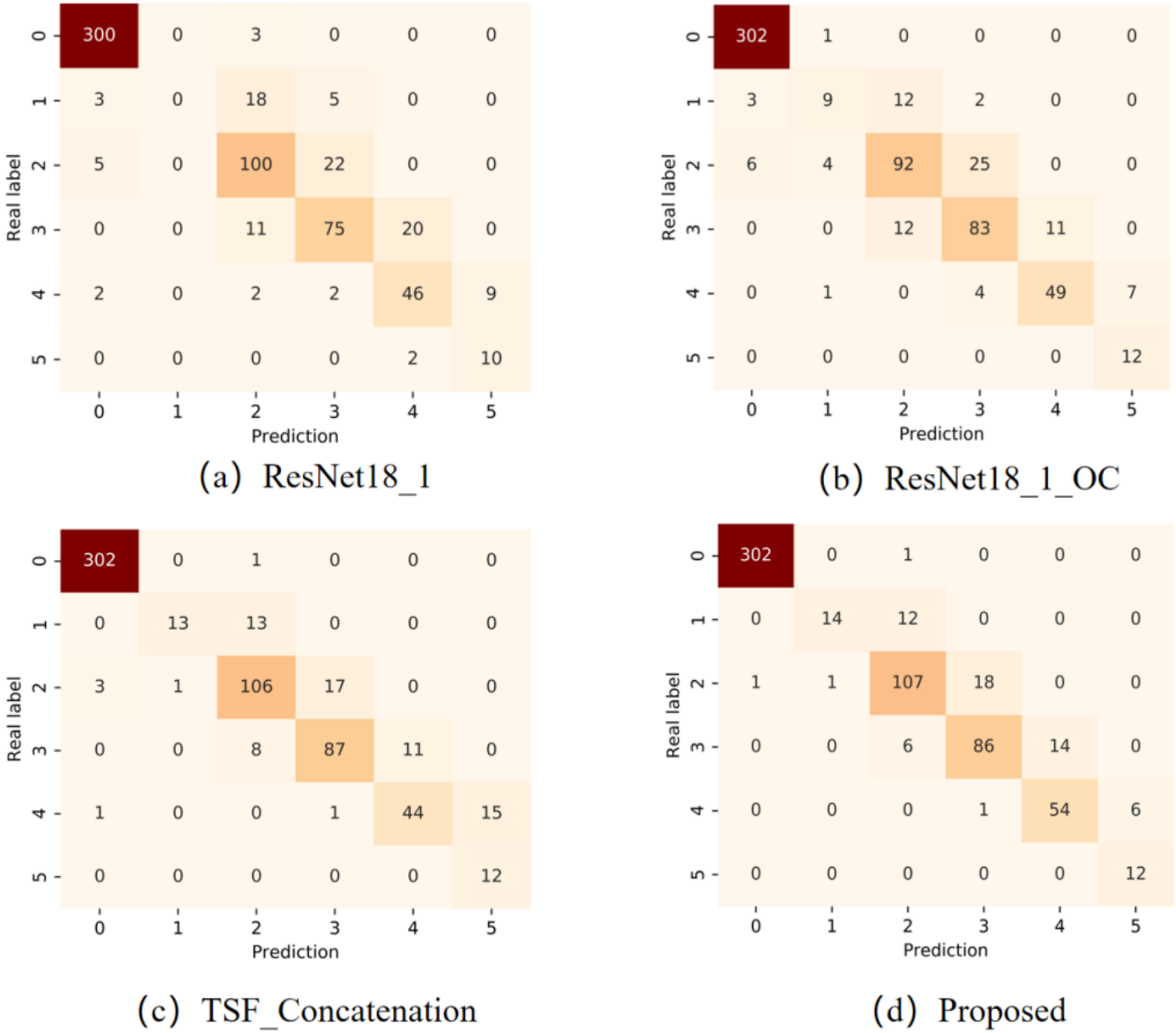


Fig. 4. The confusion matrices of different methods.

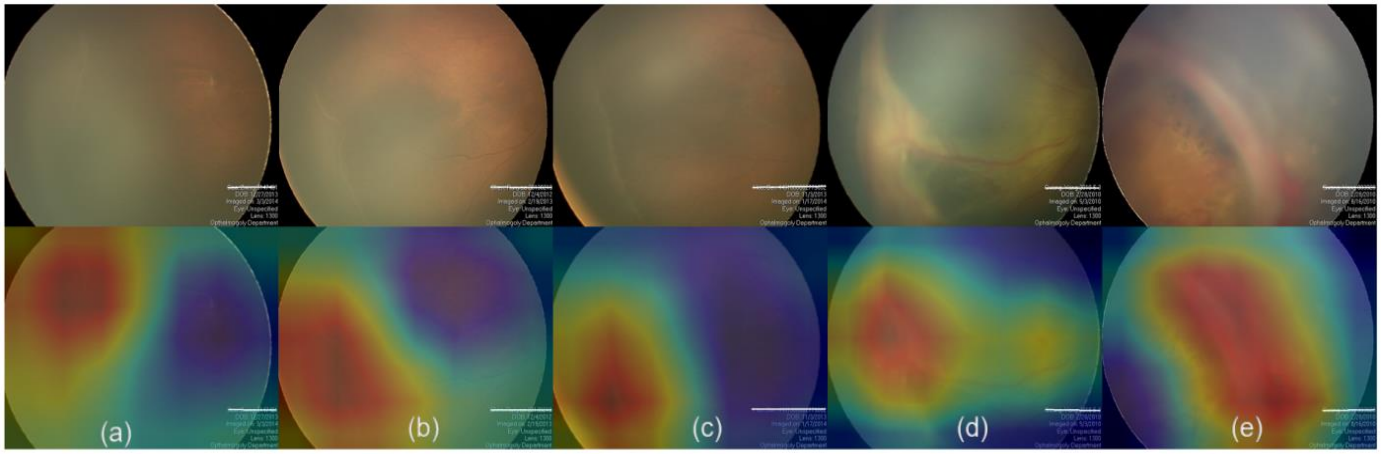


Fig. 5. Original fundus images and their corresponding heat maps of class activation. (a) Stage 1. (b) Stage 2. (c) Stage 3. (d) Stage 4. (e) Stage 5.

C. Results of the Proposed Method

1) Results of Per-image ROP staging

For per-image classification pattern, we validate the proposed method on the 635 fundus images of ROP from 196 examinations. Table IV shows the quantitative results of different methods for per-image classification pattern. As can be observed from Table IV, for per-image classification pattern, our method improves the W_R by 6.93%, 5.67% and 6.46% compared with ResNet18_1, DenseNet121_1 and EfficientNetB2_1 respectively, and achieves 0.9055 for W_R, 0.9092 for W_P, 0.9043 for W_F1, 0.9827 for ACC1 and 0.9786 for kappa. To analyze the classification performance of our method for each stage of ROP, the confusion matrices shown in Fig. 4 are used to compare the classification results of our proposed method with baseline networks. As can be seen from Fig. 4, the proposed model can identify each stage of ROP with high accuracy except for stage 1, which outperforms the ResNet18_1 and ResNet18_1_OC in each index. We introduce the "class activation mapping" technology proposed by Zhou et al [60] to obtain the heat maps of fundus images with different ROP stages, which are shown in Fig. 5. As shown in Fig. 5, the heat maps can focus on the location of key lesions related to different ROP stages, which demonstrates the effectiveness of the proposed method in the ROP staging.

2) Results of Per-examination ROP Staging

For per-examination classification pattern, we evaluate the performance of the proposed method with a 4-fold cross validation strategy. Table V shows the quantitative results of different methods for per-examination classification pattern. As can be seen from Table V, for per-examination classification pattern, our method improves the W_R by 5.32%, 2.82% and 4.63% compared with ResNet18_1, DenseNet121_1 and EfficientNetB2_1 respectively, and achieves 0.8006 for W_R, 0.8039 for W_P, 0.7958 for W_F1, 0.8706 for ACC1 and 0.9107 for Kappa. In addition, as can be seen from Table IV, the performance of per-image staging classifier is higher than per-examination. The possible reason is that the amount of training data for per-image ROP staging is relatively larger than that for per-examination ROP staging, which is beneficial to the model training, especially for the relatively complex deep network.

3) Comparison with Other Methods

The proposed network is compared with other state-of-art methods. First, taking the per-image classification for example, we compare our method with other excellent classification networks, including Inception-V4 [61], ResNext50 [62], SE_ResNet50 [63], SE_ResNext50 [63]. The quantitative test results are present in Table VI. As can be seen from Table VI, the proposed method outperforms other classification networks in our ROP staging task in terms of all metrics. Compared to the second best network (SE_ResNext50_OC), the W_R, W_P, W_F1, ACC1 and Kappa of the proposed method increase by 2.20%, 1.87%, 1.96%, 1.26% and 1.56%, respectively. As also can be seen from Table VI, the strategy of ordinal classification can improve the performance of ROP staging for all baseline networks, which shows its good universality.

Second, we compare our method with four recent studies on ROP analysis [19], [21], [22] and [23]. The network proposed by Hu et al. [19] is a per-examination classifier based on an ImageNet pretrained Inception-V4. As can be seen from Table VII, the W_R, W_P, W_F1, ACC1 and Kappa of the proposed method are 2.55%, 8.87%, 6.53%, 3.66% and 5.84% higher than that of Hu et al.'s method. The methods proposed by Zhang et al. [21], our previous work [22] and Lei et al. [23] are per-image classifiers, which are ImageNet pretrained VGG-16, ImageNet pretrained Resnet18 with attention mechanism and ImageNet pretrained ResNet50, respectively. As can be observed from Table VIII, for the per-image classification task, the proposed method outperforms the other three methods on all metrics, which improves the W_R by 3.31%, 6.41% and 5.36% compared with the other three methods respectively.

In conclusion, our method is superior to previous studies both in ROP staging per-examination and per-image. There are two possible reasons. First, the previous researches mainly focused on the binary classification problem, which is relatively simple. Therefore, the general deep learning classification network only using a single network feature extraction can achieve better classification performance. However, in the current study, 5-level ROP staging combined with normal fundus images involves a total of 6 categories of classification recognition, in which the distinction between categories is not obvious compared to binary classification, which leads to more difficulties. Therefore, general deep learning methods may not work well, which only use a single network feature extraction. Feature fusion technology makes it possible to combine features extracted from different feature extractors for final classification, which can improve the classification accuracy. This further shows that the multi-networks feature fusion method proposed in this paper can better extract more comprehensive and rich features. Second, ROP stage from 1 to 5 is a gradual process from mild to severe, so the introduction of ordinal classification strategy is reasonable, which can effectively improve the classification accuracy.

D. Ablation study

1) Ablation Study for Adopting Transfer Learning

We propose a three-stream framework with three different parallel deep networks, which are pre-trained on ImageNet. Previous studies [48-52] suggest that, compared with training from scratch, transfer learning can greatly help to optimize the model, speed up the training convergence and solve the problem of data imbalance and overfitting. We have also conducted experiments to compare the results based on pre-training with those from scratch. As can be seen from Fig. 6, the performance based on transfer learning outperforms that from scratch significantly. As can be seen from Table IV, the ResNet18_1 (ResNet18 pre-trained on ImageNet) achieves a better performance by adopting the transfer learning, with W_R, W_P, W_F1, ACC1 and Kappa increasing from 0.5449, 0.4775, 0.4872, 0.5906 and 0.2593 (ResNet18 trained from scratch) to 0.8362, 0.8362, 0.8063, 0.8202, 0.9449 and 0.9424,

respectively. The results indicate that transfer learning is effective in our task.

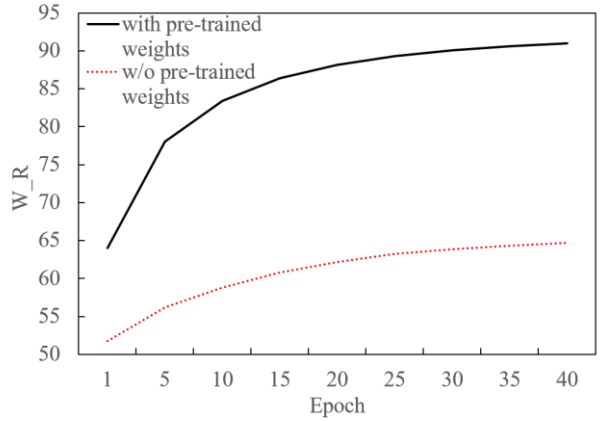


Fig. 6. Comparison between the network with pre-trained weights and the one trained from scratch.

2) Ablation Study for Ordinal Classification

To prove the effect of the ordinal classification, we conduct the ablation experiments as shown in Table IV (Row 2 to 7) and Table V (Row 1 to 6) for per-image and per-examination classification, respectively. For per-image classification, taking ResNet18_1 for example, by applying the ordinal classification (ResNet18_1_OC), the W_R, W_P, W_F1, ACC1 and Kappa increase from 0.8362, 0.8063, 0.8202, 0.9449 and 0.9424 to 0.8614, 0.8585, 0.8572, 0.9480 and 0.9610, respectively, which show that considering the order of categories can indeed improve the ROP staging performance effectively. The confusion matrices shown in Fig.4 (a) and (b) also indicate that the classifier with ordinal classification can improve the classification accuracy of most stages, especially stage 1. But the classification accuracy of stage 2 has decreased after the introduction of ordinal classification, in which four and three stage 2 images were misclassified into the adjacent stage 1 and stage 3, respectively. The possible reason is that atypical ROP fundus images with blurry boundaries are easily misclassified into adjacent stages. For per-examination classification, taking DenseNet121_1 for example, the ordinal classification can improve all quantitative indexes except ACC1 and the W_R, W_P, W_F1 and Kappa are improved by 0.81%, 6.76%, 3.83% and 3.77%, respectively, which also demonstrates ordinal classification can improve the overall performance of ROP staging.

3) Ablation Study for Different Features Fusion Strategies

In this section, we explore the influence of different fusion strategies to the ROP staging. To be consistent with our proposed three-stream parallel network structure, we use three-stream parallel network in all experiments. Here, for per-image classification, we take DenseNet121 as an example to compare four different feature fusion strategies. Table IX shows the classification results of different feature fusion strategies. As can be seen from Table IX, the feature fusion based on concatenation performs better than the other three strategies in most of the quantitative metrics. In terms of Kappa, the feature fusion based on concatenation improves by 4.35%, 0.56% and 1.60% respectively compared with other three feature fusion methods, which may because that the increase of the number of channels by concatenation makes the description of image features richer and more comprehensive and improves the classification performance. As can be seen from Table IX, the ACC1 index of feature fusion based on mean operation is the highest. The possible reason is that the feature obtained by mean feature fusion is the average of three features, which is inclined to classify the image as its adjacent next severe stage of ROP.

TABLE IX
ABLATION STUDY OF DIFFERENT FEATURE FUSION METHODS ON ROP DATA IN THIS PAPER.

Methods	W_R	W_P	W_F1	ACC1	Kappa
DenseNet121_3_add	0.8472	0.8603	0.8373	0.9591	0.9294
DenseNet121_3_mean	0.8724	0.8752	0.8633	**0.9811**	0.9673
DenseNet121_3_max	0.8583	0.8623	0.8474	0.9559	0.9569
DenseNet121_3_concatenation	**0.8756**	**0.8857**	**0.8637**	0.9717	**0.9729**

'DenseNet121_3_add', 'DenseNet121_3_mean', 'DenseNet121_3_max' and 'DenseNet121_3_concatenation' represent three parallel DenseNet121 pre-trained on ImageNet with addition feature fusion, mean feature fusion, maximum feature fusion and concatenation feature fusion, respectively.

TABLE X
ABLATION STUDY OF EACH COMPLEXITY ON ROP DATA IN THIS PAPER

Methods	W_R	W_P	W_F1	ACC1	Kappa	parameters
ResNet50	0.8519	0.8095	0.8263	0.9181	0.9317	**25.57M**
ResNet101	0.8346	0.7915	0.8114	0.9307	0.9398	44.56M
EfficientNetB5	0.8551	0.8551	0.8551	0.9827	0.9456	28.35M
Proposed	**0.9055**	**0.9092**	**0.9043**	**0.9827**	**0.9786**	26.83M

4) Ablation Study for Different Network Combinations

To explore the performance of different network combinations in the three-stream feature extraction framework, we conduct the ablation experiments for per-image classification as shown in Table IV (Row 2, 3, 4, 8, 9,10 and 16). As can be observed from Table IV, on one hand, the performance of multiple networks is better than that of a single network (Row 2 to 4 and row 8 to 10 in Table IV). On the other hand, our proposed framework with three different deep networks is better than other three frameworks with the same network in all quantitative metrics (Row 8 to 10 and row 16 in Table IV). In terms of W_R, our proposed network improves by 2.99%, 1.1% and 2.83% respectively compared with other three network combinations and achieves 0.8866 for W_R,

0.8966 for W_P, 0.8867 for W_F1, 0.9732 for ACC1 and 0.9729 for Kappa. There are two main findings from Table IV. First, DenseNet121 as feature extractor can get better classification performance, improving 1.89% and 1.42% in term of W_R compared with ResNet18 and EfficientNetB2. The possible reason is that dense connections of DenseNet121 make full use of the information of each layer. Second, the classification performance of the proposed network is further improved using these three different networks as feature extractors, which shows that ResNet18 and EfficientB2 also play a positive role in our ROP staging task. The results demonstrate that increasing the number of feature extractors can improve the classification performance and the different types of feature extraction networks can increase the diversity of the features, which is crucial for ROP staging.

5) Ablation Study for Network with Different Complexity

Studies have shown that the performance of a network is to some extent related to the complexity of the network, and the increase in the number of parameters usually leads to better performance [34-38]. To prove that the performance improvement of our proposed network is irrelevant to the increase of network complexity, for per-image classification, we compare our network with other mainstream classification networks with similar or greater complexity, which is shown in Table X. As can be seen from Table X, the network complexity of this method is similar to that of ResNet50 and EfficientNetB5, but the performance is improved 5.36% and 5.04% respectively in terms of W_R. In addition, compared with ResNet101, the number of network parameters of the proposed network is much less, but the performance is better than ResNet101. The comparison results show that the improvement of ROP classification performance is not due to the increase of network complexity.

6) Ablation Study for the Number of Sub-networks

We also conduct a series of experiments to discuss the impact of the number of sub-networks on ROP staging task from two aspects. On one hand, we explore the effect of the number of the same sub-networks as feature extractors on classification performance. Taking the first baseline network ResNet18 as an example, the experimental results of 1, 2, 3 and 4 subnetworks are shown in Table IV (Row 2, 8, 11 and 12). On the other hand, we have also conducted the experiments of combining three different subnetworks in pairs. The experimental results are shown in Table IV (Row 13 to 15). There are three main findings from Table IV (Row 2 to 4, row 8 and row 11 to 16). First, as the number of sub-networks increases (Row 2, 8, 11 and 12 in Table IV), the classification performance increases first and then decreases, which shows that increasing the number of sub-networks can improve the accuracy of ROP staging in our task, but it is not that the more subnetworks, the better. The possible reason is that too complex networks are prone to over fitting, which leads to performance degradation. Second, for different networks as feature extractors, the increase in the number of networks will also bring performance gains from the second to fourth and eighth to sixteenth rows in Table IV. Third, as can be seen from the eighth to eleventh and thirteenth to sixteenth rows in table IV, when the number of networks is the same, the performance of different sub-networks as feature extractors is better than that of the same sub-networks. The results demonstrate that the feasibility of the proposed network.

IV. Conclusion

In this paper, a simple and effective framework based on three different deep convolutional networks fusion is proposed for 5-level ROP staging per-image and per-examination. Compared with single network, our three-stream framework can extract rich and diverse high-level features from input image and fuse them into a richer and more effective feature for ROP staging, in which Resnet18 can solve the performance degradation problem of deep convolution neural network via residual connections, Densenet121's dense connection mechanism can solve the gradient disappearance problem in deep network and enhance the feature propagation and feature reuse, and EfficientNetB2 can balance the resolution, depth and width of the network to achieve good efficiency and accuracy. Compared with other state-of-art classification networks, the proposed method adopts feature fusion and ordinal classification, which can adaptively focus on lesion-related area of ROP and effectively improve the accuracy of ROP staging and the generalization ability of model.

The ablation experiments show that ordinal classification can improve the ROP staging performance significantly and the concatenation feature-level fusion strategy can further improve the classification accuracy. As can be seen from the confusion matrices in Fig. 4, for the misclassified samples, our method tends to misclassify them into the adjacent next severe stage of ROP for per-image classification and similar phenomena appear in ROP staging per-examination, which is consistent with the clinical staging rules for ambiguous samples. So as can be seen from Table IV, the ACC1 index of our proposed network achieves 0.9827 for per-image classification, which is relatively higher compared with other quantitative metrics. The possible reason is that the proposed network tends to classify ROP image into the adjacent next severe stage in the misclassification cases, which does not affect the ACC1 index because ACC1 allows a wider range of outputs as “right”.

As can be seen from Fig. 4, although compared with other classification methods, our proposed method has higher recognition accuracy for stage 1, its total performance for stage 1 is relatively low compared with the performances for other stages. In terms of test data in our study, only 14 of the 26 ROP fundus images with stage 1 were correctly identified, and the remaining 12 were all mistakenly divided into stage 2. There are two main possible reasons: (1) There are relatively few ROP data of stage 1 in the training process. Although we can reduce the impact of imbalance categories by changing the weight of loss function and applying transfer learning, we still can't solve the problem caused by data imbalance essentially. (2) Both the clinical criteria for stage 1 and stage 2 of ROP are the demarcation lines, which separate the vascularized and vascularized areas. The demarcation line in stage 2 is wider than

that in stage 1. However, for some ROP fundus images, the width of the demarcation line is not easy to be distinguished for both manual labeling and automatic ROP staging. Therefore, it may be difficult for the depth network to learn the subtle differences between them, leading to prediction errors. Although there are relatively few ROP data of stage 5 in the training process, the recognition accuracy of ROP in stage 5 is quite high. As far as our test data are concerned, all 12 ROP fundus images in stage 5 can be predicted correctly by the proposed method. The possible reason is that the retina in stage 5 is completely detached, which is relatively obvious in fundus image. Therefore, our depth network can accurately learn its effective features.

In the near future, we will collect more data to build a larger and more comprehensive ROP database with relatively balanced categories and will extend the proposed method to analyze AP-ROP, plus disease and three zones of ROP, aiming to comprehensively assist the ophthalmologist in clinical diagnosis and treatment of ROP.

REFERENCES

[1] J. Chen and L. E. H. Smith. “Retinopathy of prematurity,” *Angiogenesis*, vol. 10, no. 2, pp. 133-140, 2007.

[2] S. J. Kim et al. "Retinopathy of prematurity: a review of risk factors and their clinical significance," *Survey of Ophthalmolog*, vol. 63, no.5, pp. 618-637, 2018.

[3] A. Hellström, L. E. H. Smith, and O. Dammann. "Retinopathy of prematurity," *The lancet*, vol. 382, no. 9902, pp. 1445-1457, 2013.

[4] C. Fernando et al. "How many low birthweight babies in low-and middle-income countries are preterm?" *Revista De Saúde Pública*, vol.45, no.3 pp. 607-616, 2011.

[5] C. Gilbert, J. Rahi, M. Eckstein, J. O'Sullivan, and A. Foster. "Retinopathy of prematurity in middle-income countries," *The Lancet*, vol. 350, no. 9070, pp. 12-14, 1997.

[6] T. Wu, L. Zhang, Y. Tong, Y. Qu, B. Xia, and D. Mu. “Retinopathy of prematurity among very low-birth-weight infants in China: incidence and perinatal risk factors,” *Investigative Ophthalmology & Visual Science*, vol. 59, no.2, pp.757-763, 2018.

[7] L. Visser, R. Singh, M. Young, H. Lewis, and N. McKerrow. "Guideline for the prevention, screening and treatment of retinopathy of prematurity (ROP)," *SAMJ: South African Medical Journal*, vol. 103, no.2, pp. 116-125, 2013.

[8] 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.

[9] 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.

[10] International Committee for the Classification of Retinopathy of Prematurity. "The international classification of retinopathy of prematurity revisited," *Archives of Ophthalmology*, vol. 123, no. 7, pp. 991-999, 2005.

[11] C. Wu, R. A. Petersen, and D. K. VanderVeen. "RetCam imaging for retinopathy of prematurity screening," *Journal of American Association for Pediatric Ophthalmology and Strabismus*, vol. 10, no. 2, pp. 107-111, 2006.

[12] J. Rao et al. "Trend and risk factors of low birth weight and macrosomia in south China, 2005–2017: a retrospective observational study," *Scientific Reports.* vol. 8, no. 1, pp. 1-8, 2018.

[13] 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*, vol. 139, no. 1, pp. 30-38, 2017.

[14] A. Gschließer et al. “Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity,” *American Journal of Ophthalmology*, vol. 160, no. 3, pp. 553-560, 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," *Archives of Ophthalmology,* vol. 125, no. 7, pp. 875-880, 2007.

[16] J. Peter et al. "Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis," *JAMA Ophthalmology*, vol.134, no.6, pp.651-657, 2016.

[17] D. K. Wallace, Z. Zhao, and S. F. Freedman. "A pilot study using “ROPtool” to quantify plus disease in retinopathy of prematurity," *Journal of American Association for Pediatric Ophthalmology and Strabismus,* vol. 11, no. 4, pp.381-387, 2007.

[18] E. Ataer-Cansizoglu et al. "Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the “i-ROP” system and image features associated with expert diagnosis," *Translational Vision Science & Technology,* vol. 4, no. 6, pp. 1-12, 2015.

[19] 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, 2019.

[20] J. Wang, et al. "Automated retinopathy of prematurity screening using deep neural networks," *EBioMedicine,* vol. 35, no. 1, pp.361-368, 2018.

[21] Y. Zhang et al. "Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images," *IEEE Access*, vol. 7, no. 1, pp. 10232-10241, 2018.

[22] 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, pp. 1131321-1131327.

[23] 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 International Workshop on Ophthalmic Medical Image Analysis*, Springer, 2019, pp. 165-172.

[24] 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, pp. 173-181.

[25] V. Gulshan, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA,* vol. 316, no. 22, pp.2402-2410, 2016.

[26] S. Wang, et al. "Diabetic Retinopathy Diagnosis Using Multichannel Generative Adversarial Network with Semisupervision," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp.1-12, 2020.

[27] Yifan et al. "DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 565-575, 2019.

[28] H. Liu, D. W. K. Wong, H. Fu, Y. Xuet, and J. Liu. "DeepAMD: detect early age-related macular degeneration by applying deep learning in a multiple instance learning framework," *Asian Conference on Computer Vision*. Cham, Springer, 2018, pp. 625-640.

[29] H. Wang and Y. Yu. “Deep Feature Fusion for High-Resolution Aerial Scene Classification,” *Neural Processing Letters*, vol. 51, no. 1, pp. 853-865, 2020.

[30] R. M. Anwer et al. "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 138, no.1, pp. 74-85, 2018.

[31] X. Bian, C. Chen, Y. Sheng, Y. Xu, and Q. Du. "Fusing two convolutional neural networks for high-resolution scene classification," *In 2017 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2017, pp. 3242-3245.

[32] W. Song, S. Li, L. Fang, T. Lu. "Hyperspectral Image Classification with Deep Feature Fusion Network," *IEEE Transactions on Geoscience and Remote Sensing*. vol. 56, no. 6, pp. 3173-3184, 2018.

[33] Y. Yu, and F. Liu. "A two-stream deep fusion framework for high-resolution aerial scene classification," *Computational Intelligence and Neuroscience,* 2018, pp. 1-13.

[34] C. Feichtenhofer, A. Pinz, and A. Zisserman. "Convolutional two-stream network fusion for video action recognition," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016, pp. 1933-1941.

[35] S. Chaib, H. Liu, Y. Gu, and H. Yao. "Deep feature fusion for VHR remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing.* vol. 55, no. 8, pp. 4775-4784, 2017.

[36] X. Bian, C. Chen, L. Tian, and Q. Du. "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics*

in Applied Earth Observations and Remote Sensing, vol. 10, no. 6, pp. 2889-2901, 2017.

[37] X. Lu, W. Ji, X. Li, and X. Zheng. "Bidirectional adaptive feature fusion for remote sensing scene classification," *Neurocomputing*, vol. 328, no.1, pp. 135-146, 2019.

[38] J. Cheng, Z. Wang, and G. Pollastri. "A neural network approach to ordinal regression," *In 2008 IEEE International Joint Conference on Neural Networks,* IEEE, 2008, pp. 1279-1284.

[39] E. Frank and M. Hall. "A simple approach to ordinal classification," In *European Conference on Machine Learning*, Springer, 2001, pp. 145-156.

[40] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. "Ordinal regression with multiple output CNN for age estimation," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016, pp. 4920-4928.

[41] L. Gaudette and N. Japkowicz. "Evaluation methods for ordinal classification," *In Canadian Conference on Artificial Intelligence*, Springer, 2009, pp. 207-210.

[42] L. Li and H. –T. Lin. "Ordinal regression by extended binary classification," *In Advances in Neural Information Processing Systems*, 2007, pp. 865-872.

[43] M. Dorado-Moreno, P. A. Gutiérrez, and C. Hervás-Martínez. "Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions," *In Proceedings of the 7th international conference on Hybrid Artificial Intelligent Systems - Volume Part II* Springer, Berlin, Heidelberg, 2012, pp. 319-330.

[44] R. Potharst and J.C. Bioch. "Decision trees for ordinal classification," *Intelligent Data Analysis,* vol. 4, no. 2, pp. 97-111, 2000.

[45] 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, pp. 770-778.

[46] G. Huang et al. "Densely connected convolutional networks," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2017, pp. 4700-4708.

[47] M. Tan and Q. V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv*, *preprint arXiv:1905.11946*, 2019.

[48] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," *arXiv*, *preprint arXiv:1409.1556,* 2014.

[49] C. Szegedy et al. "Going deeper with convolutions," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015, pp. 1-9.

[50] S. J. Pan and Q. Yang. "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2009.

[51] C.Tan, F. Sun,T. Kong, W. Zhang, C. Yang, and C. Liu. "A survey on deep transfer learning," *International Conference on Artificial Neural Networks,* Springer, 2018, pp. 270-279.

[52] S. J. Pan and Q. A. Yang. "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no.10, pp. 1345-1359, 2009.

[53] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum. "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Transactions on Medical Imaging*, vol. 8, no. 3, pp. 263-269, 1989.

[54] A. D. Hoover, K. Valentina, and M. Goldbaum. "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203-210, 2000.

[55] L. Gang, O. Chutatape, and S.M. Krishnan. "Detection and measurement of retinal vessels in fundus images using amplitude modified second-order Gaussian filter," *IEEE transactions on Biomedical Engineering,* vol. 49, no. 2, pp.168-172, 2002.

[56] T. Chanwimaluang, and G. Fan. "An efficient algorithm for extraction of anatomical structures in retinal images," *In Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429),* IEEE, 2003, pp. I-1093.

[57] X. Jiang, and M. Daniel. "Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 1, pp.131-137, 2003.

[58] J. Carletta. "Assessing agreement on classification tasks: the kappa statistic," *arXiv preprint cmp-lg/9602004* ,1996.

[59] M. L. McHugh. "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276-282, 2012.

[60] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning deep features for discriminative localization," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016, pp. 2921-2929.

[61] C. Szegedy, S. Loffe, V. Vincent, and A. Alex. "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[62] 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, pp. 1492-1500.

[63] J. Hu, S. Li, and G. Sun. "Squeeze-and-excitation networks," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2018, pp.7132-7141.