tvst

Artificial Intelligence

Enhancing Retinal Fundus Image Quality Assessment With Swin-Transformer–Based Learning Across Multiple Color-Spaces

Chengcheng Huang[1,*], Yukang Jiang[2,*], Xiaochun Yang[3,*], Chiyu Wei[1], Hongyu Chen[4], Weixue Xiong[1], Henghui Lin[1], Xueqin Wang[5], Ting Tian[2], and Haizhu Tan[1]

[1] Department of Preventive Medicine, Shantou University Medical College, Shantou, China
[2] School of Mathematics, Sun Yat-Sen University, Guangzhou, Guangdong, China
[3] The First People's Hospital of Yun Nan Province, Kunming, China
[4] Department of Optoelectronic Information Science and Engineering, Physical and Materials Science College, Guangzhou University, Guangzhou, China
[5] School of Management, University of Science and Technology of China, Hefei, Anhui, China

Correspondence: Haizhu Tan, Department of Preventive Medicine, Shantou University Medical College, Shantou 515000, China. e-mail: linnanqia@126.com
Ting Tian, School of Mathematics, Sun Yat-Sen University, 135 Xingang Xi Road, Guangzhou, Guangdong 510275, China. e-mail: tiant55@mail.sysu.edu.cn

Received: June 8, 2023
Accepted: February 18, 2024
Published: April 3, 2024

Keywords: retinal image quality assessment; multiple color space; swin-transformer; score-CAM

Citation: Huang C, Jiang Y, Yang X, Wei C, Chen H, Xiong W, Lin H, Wang X, Tian T, Tan H. Enhancing retinal fundus image quality assessment with swin-transformer–based learning across multiple color-spaces. Transl Vis Sci Technol. 2024;13(4):8, https://doi.org/10.1167/tvst.13.4.8

Purpose: The assessment of retinal image (RI) quality holds significant importance in both clinical trials and large datasets, because suboptimal images can potentially conceal early signs of diseases, thereby resulting in inaccurate medical diagnoses. This study aims to develop an automatic method for Retinal Image Quality Assessment (RIQA) that incorporates visual explanations, aiming to comprehensively evaluate the quality of retinal fundus images (RIs).

Methods: We developed an automatic RIQA system, named Swin-MCSFNet, utilizing 28,792 RIs from the EyeQ dataset, as well as 2000 images from the EyePACS dataset and an additional 1,000 images from the OIA-ODIR dataset. After preprocessing, including cropping black regions, data augmentation, and normalization, a Swin-MCSFNet classifier based on the Swin-Transformer for multiple color-space fusion was proposed to grade the quality of RIs. The generalizability of Swin-MCSFNet was validated across multiple data centers. Additionally, for enhanced interpretability, a Score-CAM–generated heatmap was applied to provide visual explanations.

Results: Experimental results reveal that the proposed Swin-MCSFNet achieves promising performance, yielding a micro-receiver operating characteristic (ROC) of 0.93 and ROC scores of 0.96, 0.81, and 0.96 for the "Good," "Usable," and "Reject" categories, respectively. These scores underscore the accuracy of RIQA based on Swin-MCSF in distinguishing among the three categories. Furthermore, heatmaps generated across different RIQA classification scores and various color spaces suggest that regions in the retinal images from multiple color spaces contribute significantly to the decision-making process of the Swin-MCSFNet classifier.

Conclusions: Our study demonstrates that the proposed Swin-MCSFNet outperforms other methods in experiments conducted on multiple datasets, as evidenced by the superior performance metrics and insightful Score-CAM heatmaps.

Translational Relevance: This study constructs a new retinal image quality evaluation system, which will contribute to the subsequent research of retinal images.

Introduction

Retinal image analysis is critical for identifying and classifying various retinal diseases, including diabetic retinopathy (DR), age-related macular degeneration (AMD), retinoblastoma, hypertensive retinopathy, and retinitis pigmentosa. The automated identification of fundus diseases from retinal images represents a crucial stride toward early diagnosis and the prevention of

disease progression.[1] Retinal image quality assessment (RIQA) emerges as a pivotal prerequisite for diagnosing retinal diseases, aiming to present clear depictions of anatomical structures and lesions that are of utmost concern to ophthalmologists, while concurrently excluding poor-quality retinal images.[2] The noninvasive and cost-effective nature of retinal image acquisition on a large scale underscores its significance.[3,4] However, the quality of some retinal images is compromised because of low contrast, blurring, and focusing errors, potentially obscuring early disease signs and leading to unreliable medical diagnoses.[5] A study based on UK Biobank[6] reported that 26% of retinal images lacked adequate quality, hindering accurate diagnoses. Another study indicated that 10% and 20.8% of images with dilated and nondilated pupils, respectively, were unsuitable for Automatic Retinal Screening Systems.[7] Approximately 10% to 15% of retinal images are rejected because of poor image quality.[8] Consequently, RIQA is indispensable to circumvent the need for image recapture and ensure sufficient quality for reliable diagnoses.

In recent years, various methods have been proposed for retinal image quality classification. Although some approaches, such as those by Dias et al.,[9] Lee et al.,[10] and Abdel-Hamid et al.,[11] rely on hand-crafted features and lack generalization ability, deep learning (DL) for feature learning has emerged as a robust alternative. This approach unveils latent feature information, enabling the construction of end-to-end models for tasks with heightened robustness and accuracy.[12,13] Over the past decade, Convolutional Neural Networks (CNNs) have demonstrated remarkable success in the classification of retinal images.[14] For instance, Zago et al.[15] proposed a CNN pretrained on nonmedical data, evaluating its performance on publicly available databases (DRIMDB and ELSA-Brasil). FengLi et al.[16] introduced a method that combines saliency maps and CNN features, subsequently feeding them into a support vector machine for automated detection of retinal fundus images of varying quality. Sun et al.[17] applied two fine-tuned CNN architectures, achieving an impressive accuracy of 97.12%. Despite the achievements of CNNs, visual transformers have shown superior capabilities in capturing spatial relationships compared to CNNs. Transformer networks, including initial and visual transformers based on self-attention mechanisms,[18] have been proposed to grade retinal fundus images with RGB,[19,20] owing to their capacity to capture long-distance dependencies.

The representation of colors in retinal fundus images through color spaces such as RGB, CIELAB, and HSV, obtained via nonlinear conversions from the RGB color space, is a critical aspect. However, certain existing RIQA methods based on initial transformers and visual transformers need refinement to accommodate retinal fundus images with multiple color spaces. This is attributed to their exclusive focus on RGB, overlooking other color spaces and impeding the extraction of diverse visual features. Our choice of the Swin Transformer as the backbone network stems from its enhanced speed, specialized architecture, efficient Winograd convolutions, and the ability to preserve spatial relationships, thereby enabling more accurate image classification.

Furthermore, visualization techniques, including Class Activation Map (CAM),[21] Grad-CAM,[22] Grad-CAM++,[23] Score-CAM,[24] and Layer-CAM,[25] and more, have enabled us to visualize the predicted class scores and discerning object parts detected by CNNs. The novel post-hoc visual explanation method, Score-CAM, addresses issues such as gradient saturation, disappearance, and noise, ensuring accurate visualization. It overcomes challenges posed by channels with heavy weights in CAM and Grad-CAM, contributing disproportionately to category prediction scores. Furthermore, it addresses the limitations of CAM and Grad-CAM by considering the impact of forward prediction on backpropagation gradients.

This study delves into the analysis of different color spaces' impact on retinal images in RIQA tasks. We propose Swin-MCSFNet, a multicolor space fusion method based on the Swin Transformer, to integrate representations from various color spaces. Additionally, we validate our approach through multi-center assessments using external datasets, OIA-ODIR, and EyePACS. For enhanced visual interpretation, we use the visualization technology Score-CAM.

Material and Methods

The comprehensive workflow is illustrated in Figure 1 and is detailed in the subsequent subsections.

Dataset

Retinal fundus images from multiple centers, including EyePACS, Eye-Quality (EyeQ), and OIA-ODIR, were collected in our study. EyePACS, a versatile telemedicine system dedicated to diabetic retinopathy screening, has collected five million retinal images from more than 750,000 screened patients.[26] The EyeQ Assessment Dataset, a re-annotation subset of the EyePACS dataset, is specifically curated for RIQA.[3]

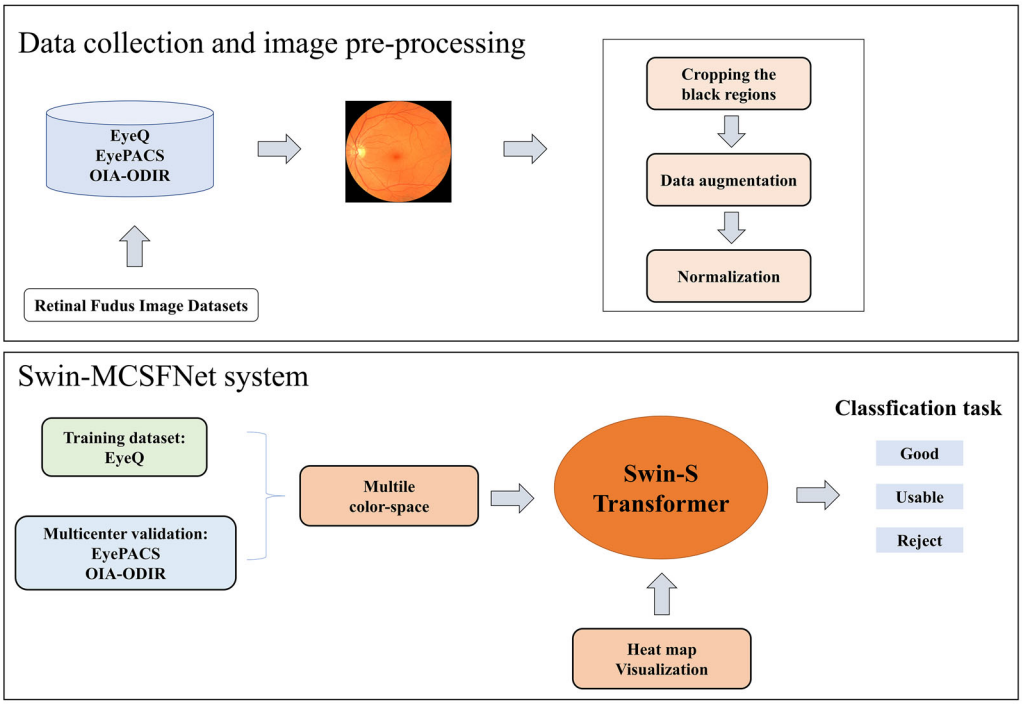

Figure 1. The entire workflow of the proposed Swin-MCSFNet.

The inherent heterogeneity in image characteristics across multiple centers can be ascribed to the use of more than one camera model and various types of cameras. This diversity introduces variations in resolutions, fields-of-view, hues, retinal image centers, pupil diameters, and other relevant factors.

Some previous studies[27–29] have demonstrated a considerable number of retinal fundus images, previously deemed suboptimal (21%) by certain two-level RIQA systems (categorizing images as "good" or "bad"), could still offer valuable interpretability for clinicians.[30] To address this issue, EyeQ dataset and Fu et al.[3] proposed a three-level RIQA system respectively. In our study, considering the quality of the collected retinal fundus images, we opted for the adoption of a three-level RIQA system. This system, named Swin-MCSFNet, uses a Swin-Transformer for multiple color-space fusion, deviating from the conventional two-level RIQA system to enhance clinical relevance. Our three-level RIQA system is defined as follows: "Good" quality denotes retinal images devoid of low-quality factors, displaying all retinopathy characteristics clearly. "Usable" quality includes retinal images with slight low-quality indicators, where main structures and lesions are still clearly identifiable by the graders. In instances of uneven illumination, where the readable region of the retinal image exceeds 80%, these images are considered "Usable." "Reject" grade is assigned to images with severe issues preventing a full and reliable diagnosis, even by ophthalmologists, including those with an invisible disc or macula region.

Therefore the retinal fundus images gathered from the EyeQ dataset in our study underwent grading, classifying them into "good," "usable," and "reject" categories. Within this classification, 12,543 images were assigned to the training dataset, 4234 to the validation dataset, and 12,015 to the testing dataset (refer to Table 1). Additionally, in accordance with the quality grading system of the EyeQ dataset, 1000 retinal fundus images from the OIA-ODIR and 2000 selected images from the EyePACS dataset were systematically categorized into the same three-level quality grading system. This categorization was validated using the EyeQ dataset, OIA-ODIR datasets by three specialized retinal care graders at the First People's Hospital of Yun Nan Province.

Preprocessing on the Dataset

Cropping the Black Regions From the Images

To ensure the uniformity of network input and eliminate irrelevant information, we conducted preprocessing on the obtained images, specifically by cropping the black regions of the retinal fundus images, which were surrounded by black margins. Subsequently, the images were resized to 256 × 256 pixels. Figure 2 illustrates a sample from EyePACS, depicting the image before and after the removal of borders.

Table 1. A Description of Each Dataset

Dataset	Total	Training Set	Validation Set	Test Set
EyeQ	28,792	12,543 (Good: 8347, Usable: 1876, Reject: 2320)	4234	12,015 (Good: 6277, Usable: 3390, Reject: 2348)
OIA-ODIR	1000	—	—	1000 (Good: 538, Usable: 280, Reject: 182)
EyePACS	2000	—	—	2000 (Good: 620, Usable: 1146, Reject: 234)

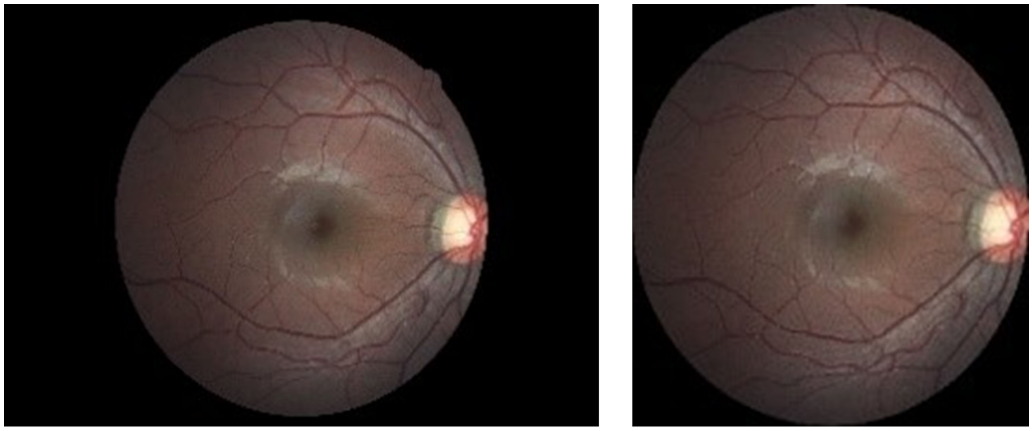

Figure 2. Sample image from EyePACS before and after cropping, (**a**) one sample image with black margins, (**b**) cropped and centered image.

Data Augmentation

The application of data augmentation on the original data has been demonstrated to enhance diagnostic accuracy in previous studies.[31–34] In our study, two types of augmentation were used:

1. Randomly inverting images vertically and horizontally with a probability of 0.5
2. Performing affine transformations through rotation with random angles ranging from −180° to 180°.

It is important to note that data augmentation was not applied to the testing dataset to ensure the validity of the evaluation results.

Normalization

Before inputting images from all used datasets into the deep learning network, a normalization process was implemented. This involved calculating the global mean and standard deviation (SD) of pixel values across all training images.

Swin-MCSFNet Classifier: Image Quality Classifier Based on Swin Transformer for Multiple Color-Space Fusion

In our study, we introduced the Swin-transformer-based Multiple Color-space Fusion network (Swin-MCSFNet) to integrate representations from various color spaces. The architecture of the proposed Swin-MCSFNet classifier is illustrated in Figure 3.

The original RGB retinal fundus image underwent color image segmentation, using the HSV and LAB color spaces.[35] Subsequently, a two-dimensional convolution patch partition module was used to segment the images into non-overlapping patches of size 4×4. This resulted in a feature dimension of $4 \times 4 \times 3$ for each patch. The patches were then flattened in the channel dimension, producing a $(\frac{H}{4} \times \frac{W}{4}) \times 48$ two-dimensional sequence. A Linear Embedding layer was used to map the tensor with dimensions $(\frac{H}{4} \times \frac{W}{4}) \times 48$ to a dimension "C" (set to 96), yielding the resultant patch tokens. These tokens were then input into the Base Networks, where image features were generated through the

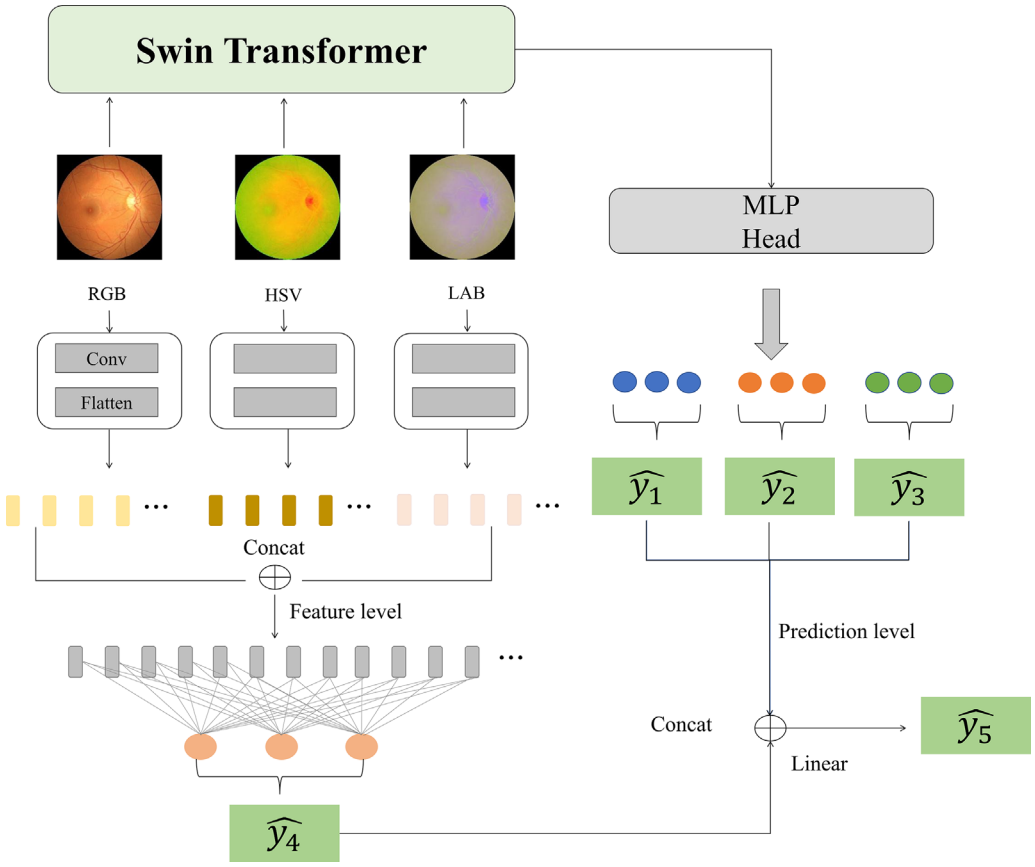

Figure 3. The archicture of the Swin-MCSFNet classifier.

use of the Swin-S transformer on patch tokens. The architecture of the Swin-S transformer is depicted in Figure 4a. Multiple successive Swin Transformer Blocks with W-MSA and SW-MSA head self-attention modules (Fig. 4b) were applied to these patch tokens.

Furthermore, the two-level fusion block was developed, encompassing prediction-level fusion and feature-level fusion: (1) The feature maps from the Base Networks and different color spaces were merged and fed into a fully connected layer to generate prediction-level fusion. (2) Considering the significance of detailed information for shallow networks, the patch partition module and linear embedding were used to partition patch tokens acquired from different color spaces. These tokens were then input into the adaptive average pooling layer, followed by flattening and direct feeding into the fully-connected layer for feature-level fusion. Ultimately, the final prediction-level fusion was achieved by combining the outputs of the two-level fusion block and feeding them into the fully connected layer. This two-level fusion block ensured comprehensive use of information from different color spaces, thereby enhancing the overall system's accuracy.

In addition, the proposed Swin-MCSFNet retained the loss function of all three networks and shallow networks through the multi-branch fusion network, combined with the fusion loss as:

$$Loss_{total} = \sum_{i=1}^{3} w_i Loss_i + w_f Loss_f + w_p Loss_p \quad (1)$$

where, w_i, w_f and w_f were tradeoff weights, which were set as $w_i = 0.1$, $w_f = 0.1$, and $w_f = 0.6$[3].To emphasize the final prediction fusion layer, w_p is to 0.6. $Loss_i$, $Loss_f$ and $Loss_p$were the multiclass cross entropy loss functions of the three base networks and two fusion layers, respectively.

Quality Inspection of Retinal Images with Multiple Color Spaces Using Score-CAM

Score-CAM was introduced to conduct quality inspections of fundus color photos in multiple color spaces, with the aim of achieving a qualitative analysis of various attention modules (see Fig. 5).

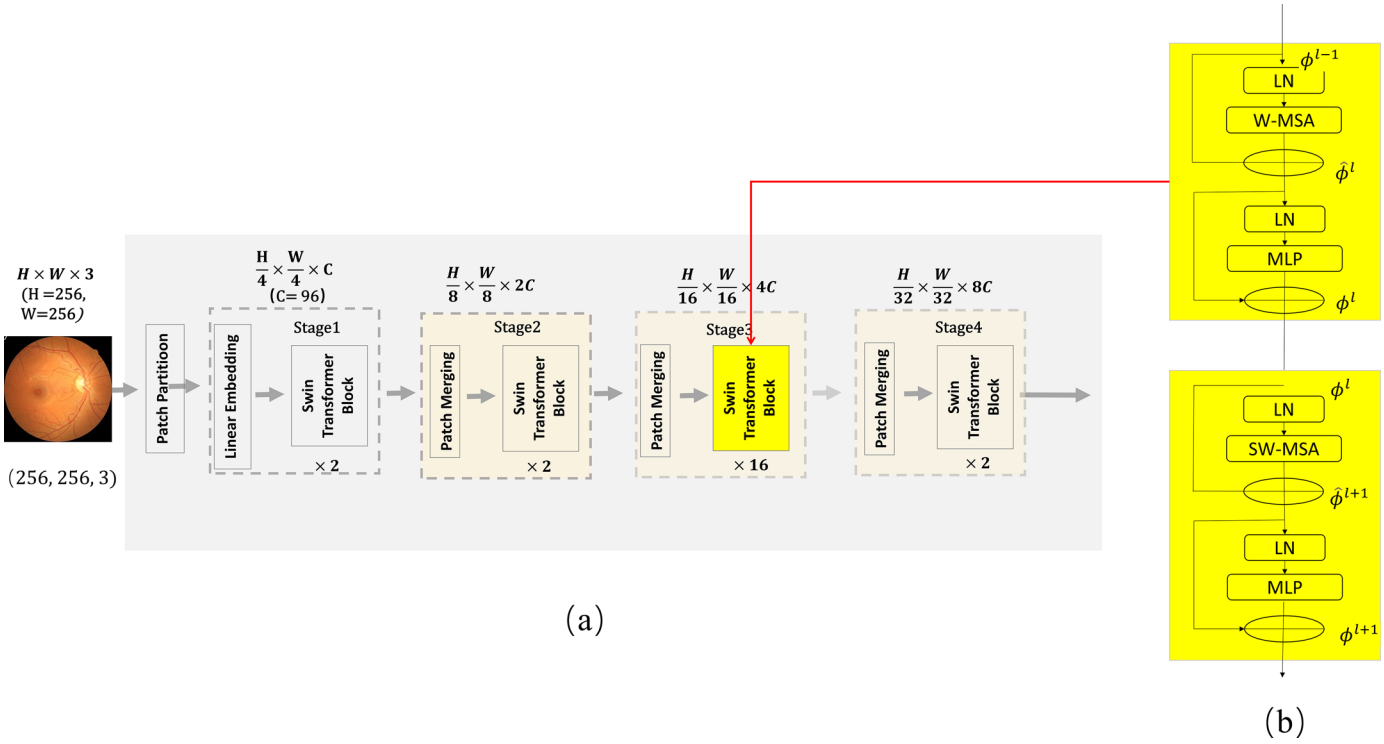

Figure 4. (**a**) The architecture of a Swin transformer (Swin-S), (**b**) Multiple successive Swin Transformer Block with multi-head self-attention modules.

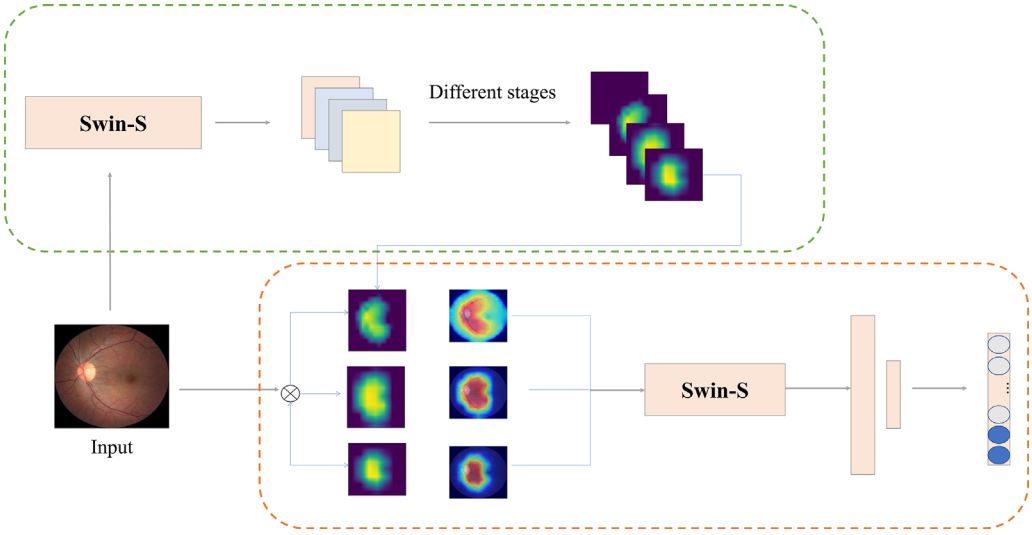

Figure 5. Use Score-CAM to visualize different stages of Swin-s on retinal images.

Furthermore, it is crucial to determine the specific areas of the pre-processed retinal fundus images from which the network is learning, focusing on relevant information rather than irrelevant details for classification. The heatmaps generated by Score-CAM serve the purpose of aiding in the accurate interpretation of CNN predictions with reduced noise.

Experiments

Experimental Setting

We used 12,543 retinal fundus images from the EyeQ dataset as training data, and 12,015 retinal fundus images from the same dataset were used as testing data. To assess the generalization capability of our proposed system, Swin-MCSFNet, 4234 images

from the EyeQ dataset, along with 2000 pictures from EyePACS and 2000 pictures from OIA-ODIR, were adopted. For a multicenter comparison, we evaluated the proposed Swin-MCSFNet by leveraging two state-of-the-art networks: Swin-S[36] and DenseNet121-MCS.[3] To assess the impact of the network's color space, we compared Swin-MCSFNet for each base network. The implementation of the multi-branch fusion classification network for RIQA based on Swin-MCSFNet was conducted using PyTorch. The parameters were set with a batch size of four, a learning rate of 0.01, an epoch of 100, stochastic gradient descent as the optimizer, and a multi-class cross-entropy loss function as the loss function. All experiments were conducted on a Hygon C86 3185 eight-core processor machine with 65G RAM, equipped with an NVIDIA GeForce RTX 3090 GPU (VRAM: 24G). The training of our model on the EyeQ dataset took approximately nine hours, and testing was completed within five minutes.

The loss function was calculated as follows:

$$Loss\,(y_i, p_i) = -\sum_{i=1}^{C} y_i log\,(p_i)$$

where C is the number of classes. y_i denoted the one hot value representation of the label, p_i denoted the probability of the i^{th} class.

We evaluated Swin-MCSFNet by employing metrics such as accuracy, precision, recall, and F1-score. Finally, our study used a heatmap to visually represent the performance and efficiency of the proposed RIQA based on Swin-MCSFNet. This visualization enabled us to compare the performance of different components or configurations and identify potential issues or areas for improvement.

Results

The performance of RIQA using three different methods on retinal fundus images with multiple color spaces is summarized in Tables 2 through 4. The results, encompassing metric accuracy, precision, and F1-score, highlight the superior performance of the proposed method (Swin-MCSFNet) across multiple centers compared to the other two methods. Figure 6 presents the ROC curves and confusion matrix of the three methods on the EyeQ testing dataset. The evaluation of classifying retinal fundus images into "Good," "Usable," and "Bad" categories demonstrates the effectiveness of the proposed Swin-MCSFNet RIQA method. Notably, Swin-MCSFNet achieved ROC scores of 0.96, 0.81, and 0.96 in the "Good," "Usable," and "Rejected" categories, respectively, with a micro ROC score of 0.93. These results underscore the enhanced performance of Swin-MCSFNet in distinguishing among the three categories, as evidenced by the ROC curves. The binary classification results between different models are shown in the supplementary materials. Figures 6 and 7 offer deeper insights into the classification process through the inclusion of confusion matrices and heatmap visualizations. The network's focal points are evident, primarily centering on the optic cup and disc for both "Good" and "Usable" retinal images. Conversely, retinal fundus images categorized as "Reject" demonstrate distinctive characteristics arising from issues such as dim lighting, poor contrast, and haziness. It is noteworthy that Figure 7 illustrates only two representative types of "Reject" images. The Score-CAM heatmap of these poor-quality retinal fundus images prominently exhibits a red hue, signifying a concentration of overall information in specific areas. This underscores the network's emphasis on critical regions, even in poor-quality images. Figure 8 enhances our understanding of heatmap visualizations across different color spaces. Specifically, the RGB image consistently emphasizes areas surrounding the optic cup and disc. In contrast, the HSV and LAB color spaces prioritize color intensity. Particularly noteworthy is the LAB image's heatmap, which exhibits a comprehensive focus on the entire image, encompassing even blood vessels.

Table 2. Performance Comparison of RIQA Using Different Methods on the EyeQ Training Dataset

Network	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)
Swin-S	0.8033 (0.7962, 0.8104)	0.7659 (0.7583, 0.7734)	**0.7706** (0.7631, 0.7781)	0.7633 (0.7557, 0.7709)
DenseNet121-MCS	0.8529 (0.8488, 0.8569)	0.7409 (0.7339, 0.7480)	0.7606 (0.7545, 0.7673)	0.7340 (0.7273, 0.7410)
Ours	**0.8770** (0.8729, 0.8806)	**0.7919** (0.7845, 0.7985)	0.7658 (0.7585, 0.7721)	**0.7764** (0.7689, 0.7827)

CI, confidence interval. Bold values indicate the model with the best performance in each indicator.

Table 3. Performance Comparison of RIQA Using Different Methods on the OIA-ODIR Dataset

Network	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)
Swin-S	0.7070 (0.7900, 0.8200)	0.6563 (0.6319, 0.6835)	**0.6674** (0.6452,0.6915)	**0.6532** (0.6294, 0.6780)
DenseNet121-MCS	0.7320 (0.7156, 0.7478)	0.5695 (0.5474, 0.5935)	0.6038 (0.5793,0.6300)	0.5600 (0.5358, 0.5852)
Ours	**0.7980** (0.7833, 0.8122)	**0.6385** (0.6122, 0.6615)	0.6645 (0.6375,0.6861)	0.6439 (0.6175, 0.6673)

CI, confidence interval. Bold values indicate the model with the best performance in each indicator.

Table 4. Performance Comparison of Three Methods on the EyePACS Dataset

Network	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)
Swin-S	0.5655 (0.5438, 0.5872)	0.5596 (0.5379, 0.5814)	0.6167 (0.5954, 0.6380)	0.5653 (0.5435, 0.5870)
DenseNet121-MCS	0.6520 (0.6406, 0.6639)	0.4767 (0.4586, 0.4949)	0.5744 (0.5562, 0.5916)	0.4709 (0.4525, 0.4899)
Ours	**0.7487** (0.7372, 0.7589)	**0.6162** (0.5969, 0.6355)	**0.6635** (0.6433,0.6817)	**0.6271** (0.6080, 0.6447)

CI, confidence interval. Bold values indicate the model with the best performance in each indicator.

Discussion

In this article, we presented the Swin-MCSFNet classifier, an effective RIQA system based on the Swin transformer. Swin-MCSFNet classifier is specifically designed for the precise classification of retinal images across multiple color spaces into three distinct grades. To validate the efficacy of the proposed Swin-MCSFNet classifier, comprehensive assessments were conducted across various centers. The justification for its performance was established using Score-CAM, a method used to identify the most critical areas in the images for evaluating image quality.

Certain retinal images acquired from the EyePACS or EyeQ dataset, initially classified as Reject quality by the implemented RIQA system, are reclassified as having Usable quality by our proposed Swin-MCSFNet classifier. The observed disparity in classification can be linked to issues such as uneven or inadequate illumination, which may result from the absence of pharmacological mydriasis or irregular operational procedures. This issue is also present in on-site eye disease screening in rural areas among community residents and college students. Figure 5 demonstrated that a considerable amount of useful information can still be extracted from retinal images of Good and Usable grades. Therefore it is essential to maximize the use of such images when available.

The study investigated the performance of three state-of-the-art RIQA classifiers, all of which yielded impressive results. As illustrated in Tables 2 through 4, the classifier built on the Swin-MCSFNet demonstrated a slight superiority over the others, as evidenced by higher accuracy, precision, F1 score, micro-ROC, and ROC scores metrics. In comparison to the classifier relying on the Swin-S transformer[36] for images with a single color space, Swin-MCSFNet exhibited the ability to use information from each color space, effectively capturing subtle differences among diverse image types and thereby enhancing prediction accuracy. A noteworthy advantage of Swin-MCSFNet over the DenseNet121-MCS-based classifier[3] lies in its capacity to simultaneously capture both global and local information, leading to improved generalization ability. Furthermore, the Swin-MCSFNet classifier offers a significant computational efficiency advantage over the DenseNet121-based classifier, enabling faster inference and training processes.

DL models exhibit a black-box architecture, posing challenges to comprehensively explore their functionality.[37] Consequently, explainable artificial intelligence models have been proposed to strike a balance between explainability and accuracy in black-box neural networks. This approach enables a deeper understanding and unveiling of the black-box behavior inherent in deep neural networks. Visualization tools, such as Score-CAM—an integral part of explainable artificial intelligence—have been developed to generate

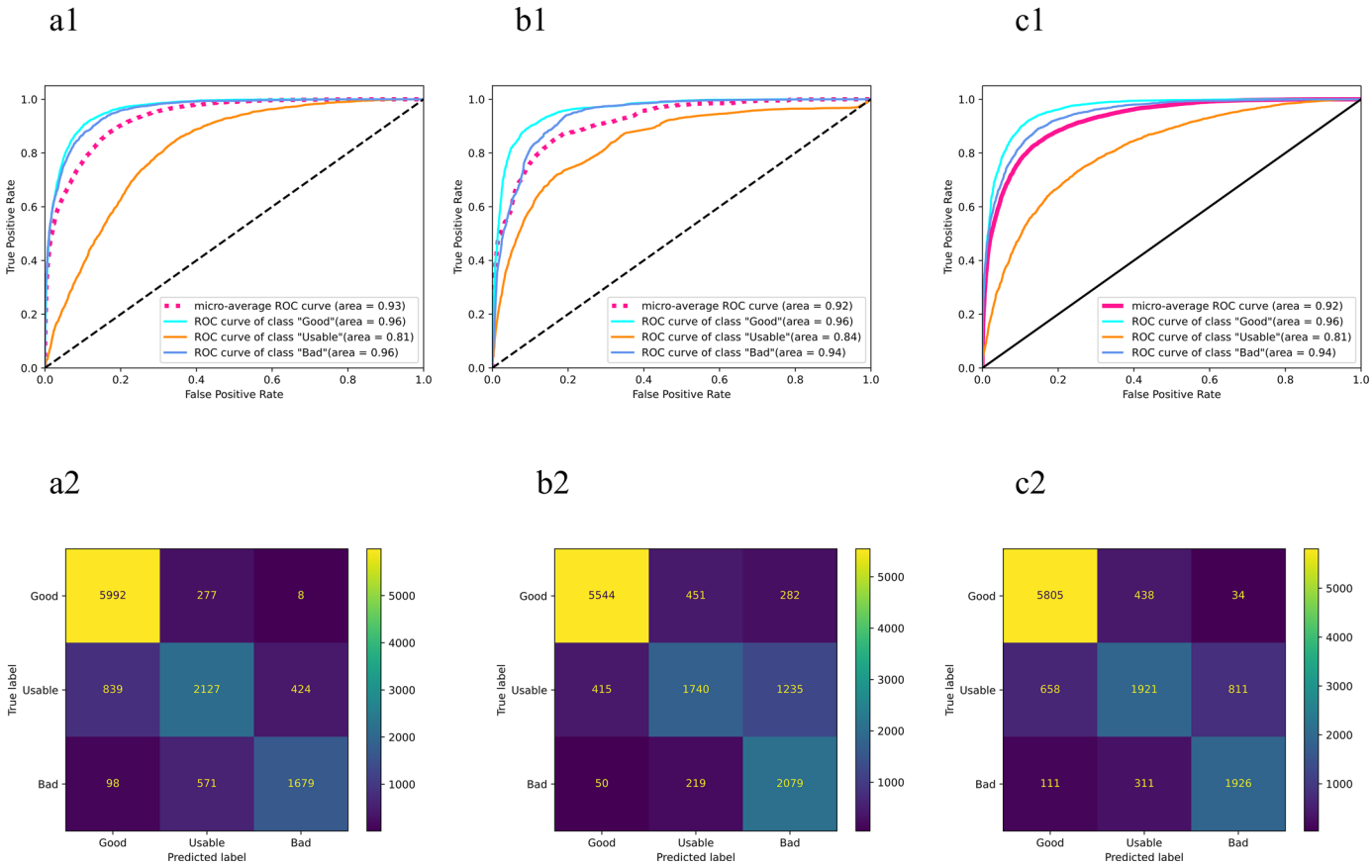

Figure 6. Comparison of ROC curve and confusion matrix of three methods on the EyeQ Dataset. (a1–a2): the proposed Swin-MCSFNet; (b1–b2): DenseNet121-MCS; (c1–c2): Swin-S

heatmaps. These heatmaps assist clinicians in quickly identifying areas that may warrant further evaluation or improvement. In our study, we used heatmaps across different RIQA classification scores (see Fig. 7) and across various color spaces (see Fig. 8). The analysis revealed that specific regions in the retinal images with multiple color spaces predominantly influenced the decision-making process of the Swin-MCSFNet classifier. This not only aids in evaluating the model's performance and understanding the decision-making process during image classification but also streamlines the quality assessment procedure. Moreover, it reduces the risk of overlooking subtle changes that might be crucial for determining the overall image quality.

Furthermore, the proposed Swin-MCSFNet classifier underwent successful validation across multiple data centers, as illustrated in Tables 2 through 4 and Figure 6. This validation serves as additional confirmation of the accuracy and reliability of the Swin-MCSFNet classifier in diverse contexts. It suggests that the proposed Swin-MCSFNet classifier is a valuable tool for both inexperienced and experienced medical professionals, enabling them to detect and grade the quality of retinal fundus images with heightened accuracy and efficiency. Additionally, our proposed classifier exhibits the capability to monitor treatment progress and identify subtle changes throughout the course of treatment. This feature contributes to earlier and more accurate diagnoses of eye diseases, facilitating faster treatments and improving outcomes for affected individuals, potentially enhancing their quality of life. Ultimately, the proposed classifier can serve as a potent tool for automating retinal screening processes. This has the potential to alleviate the workload on healthcare professionals, leading to increased efficiency in the diagnosis of retinal diseases.

One limitation of this study is the observed imbalance in sample distribution within the EyeQ dataset, as indicated in Table 1. To mitigate this issue in future research, techniques such as image enhancement or degradation could be used. Additionally, Figure 6 illustrates that the accuracy of grading in the Usable category was not notably high. This discrepancy might be attributed to the insufficient precision of the RIQA

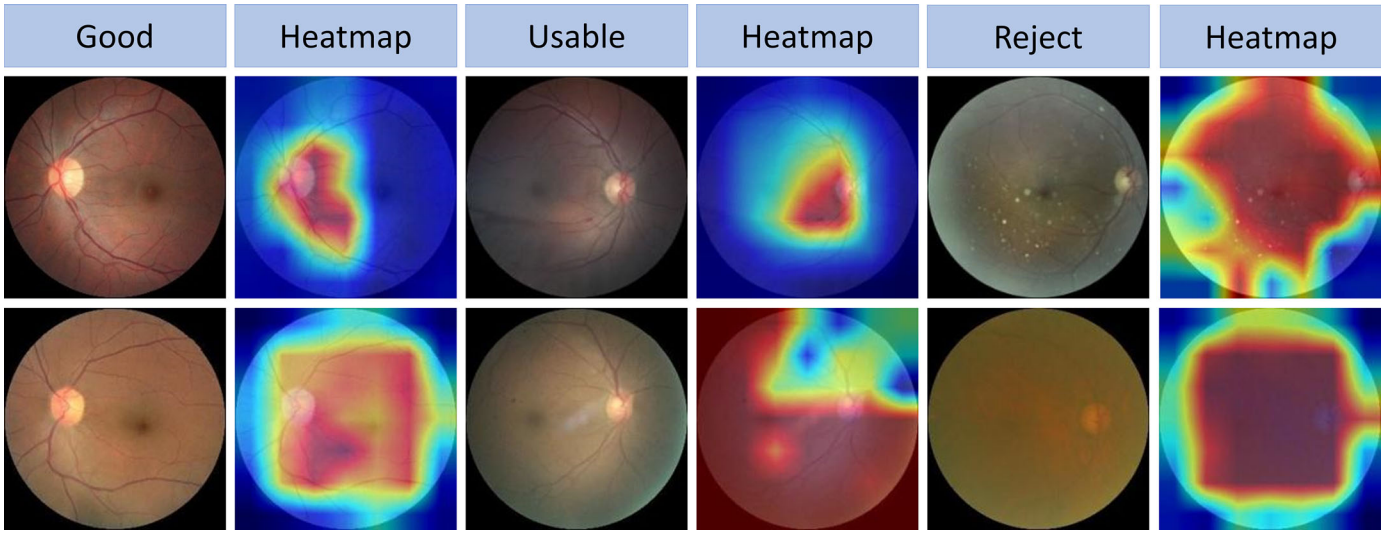

Figure 7. Heat map visualization across different RIQA classification.

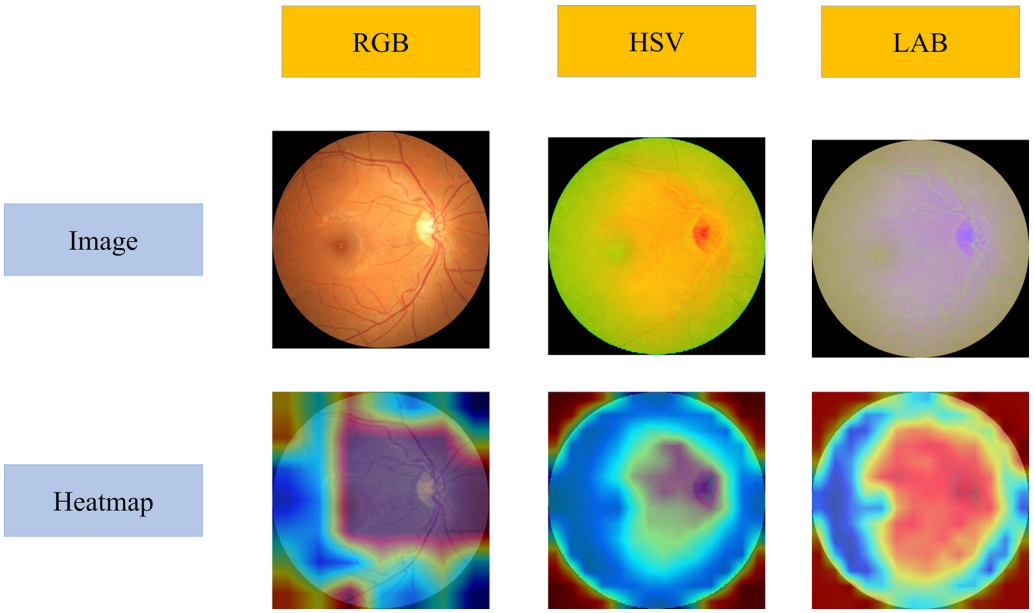

Figure 8. Heat map visualization across different color spaces.

in the EyeQ data, hindering accurate grading within the Usable categories.

Conclusions

In conclusion, our study demonstrates that the proposed Swin-MCSFNet outperformed other methods in experiments conducted across multiple datasets. Furthermore, visual explanations provided through the Score-CAM technique support the assertion that the Swin-MCSFNet classifier is an effective tool for accurately classifying retinal images. This finding holds potential long-term benefits for the medical community.

Acknowledgments

The authors thank the Department of Statistics and Finance, School of Management, and University of Science and Technology, for their guidance and use of

their research computing infrastructure. English proofreading was provided by ChatGPT.

Partially supported by the National Natural Science Foundation of China grants (No.12231017, 81770971, 12001554, and 72171216); the National Key R&D Program of China (No.2022YFA1003803), the Natural Science Foundation of Guangdong Province, China (No. 2021A1515010205 and 2020A1515010617), the Science and Technology Program of Guangzhou, China (No. 202201011578), Science and Technology Talent and Platform Plan of Yunnan Province (2019HB050 and YNWR-QNBJ-2018-315), and the Open Research Funds of the State Key Laboratory of Ophthalmology (No. 2019KF02).

Disclosure: **C. Huang**, None; **Y. Jiang**, None; **X. Yang**, None; **C. Wei**, None; **H. Chen**, None; **W. Xiong**, None; **H. Lin**, None; **X. Wang**, None; **T. Tian**, None; **H. Tan**, None

[*] CH, YJ, and XY contributed equally to this work.

References

1. Badar M, Haris M, Fatima A. Application of deep learning for retinal image analysis: a review. *Comp Sci Rev.* 2020;35:100203.
2. König M, Seeböck P, Gerendas BS, et al. Quality assessment of colour fundus and fluorescein angiography images using deep learning. *Br J Ophthalmol.* 2023;108:98–104.
3. Fu H, Wang B, Shen J, et al. Evaluation of retinal image quality assessment networks in different color-spaces. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I*. Berlin: Springer International Publishing; 2019:48–56.
4. Wang H, Meng X, Tang Q, Hao Y, Luo Y, Li J. Development and application of a standardized testset for an artificial intelligence medical device intended for the computer-aided diagnosis of diabetic retinopathy. *J Healthcare Eng.* 2023;2023.
5. Şevik U, Köse C, Berber T, Erdöl H. Identification of suitable fundus images using automated quality assessment methods. *J Biomed Opt.* Apr 2014;19(4):046006.
6. MacGillivray TJ, Cameron JR, Zhang Q, et al. Suitability of UK Biobank retinal images for automatic analysis of morphometric properties of the vasculature. *PLoS One.* 2015;10(5):e0127914.
7. Yu H, Agurto C, Barriga S, Nemeth SC, Soliz P, Zamora G. Automated image quality evaluation of retinal fundus photographs in diabetic retinopathy screening. In: 2012 IEEE Southwest symposium on image analysis and interpretation. New York: IEEE; 2012:125–128.
8. Davis H, Russell S, Barriga E, Abramoff M, Soliz P. Vision-based, real-time retinal image quality assessment. In: 2009 22nd IEEE International Symposium on Computer-Based Medical Systems. New York: IEEE; 2009.
9. Dias JMP, Oliveira CM, da Silva Cruz LA. Retinal image quality assessment using generic image quality indicators. *Information Fusion.* 2014;19:73–90.
10. Lee SC, Wang Y. Automatic retinal image quality assessment and enhancement. In: Medical imaging 1999: image processing. Cergy-Pontoise, France: SPIE; 1999;3661:1581–1590.
11. Abdel-Hamid L, El-Rafei A, El-Ramly S, Michelson G, Hornegger J. Retinal image quality assessment based on image clarity and content. *J Biomed Opt.* 2016;21(9):096007.
12. Dong C, Loy CC, He K, Tang X. Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell.* 2015;38:295–307.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:770–778.
14. Ramprasath M, Anand MV, Hariharan S. Image classification using convolutional neural networks. *Int J Pure Appl Mathematics.* 2018;119:1307–1319.
15. Zago GT, Andreao RV, Dorizzi B, Salles EOT. Retinal image quality assessment using deep learning. *Comput Biol Med.* 2018;103:64–70.
16. FengLi Y, Jing S, Annan L, Jun C, Cheng W, Jiang L. Image quality classification for DR screening using deep learning. *Annu Int Conf IEEE Eng Med Biol Soc.* Jul 2017;2017:664–667.
17. Sun J, Wan C, Cheng J, Yu F, Liu J. Retinal image quality classification using fine-tuned CNN. *Fetal, infant and ophthalmic medical image analysis*. Berlin: Springer; 2017:126–133.
18. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. *Adv Neural Inf Proc Syst.* 2021;34:15908–15919.
19. Yao Z, Yuan Y, Shi Z, et al. FunSwin: a deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. *Front Physiol.* 2022;13:961386.
20. Rodriguez M, AlMarzouqi H, Liatsis P. Multi-label retinal disease classification using transformers. *IEEE J Biomed Health Informatics.* 2022.

21. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:2921–2929.
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, 2017:618–626.
23. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV), 2018:839–847.
24. Wang H, Wang Z, Du M, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020:24–25.
25. Jiang P-T, Zhang C-B, Hou Q, Cheng M-M, Wei Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans Image Proc.* 2021;30:5875–5888.
26. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol.* 2009;3:509–516.
27. Pérez AD, Perdomo O, González FA. A lightweight deep learning model for mobile eye fundus image quality assessment. In: 15th International Symposium on Medical Information Processing and Analysis. Cergy-Pontoise, France: SPIE. 2020;11330:151–158.
28. Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated quality assessment of colour fundus images for diabetic retinopathy screening in telemedicine. *J Digit Imaging.* 2018;31:869–878.
29. Chan E, Tang Z, Najjar RP, et al. A deep learning system for automated quality evaluation of optic disc photographs in neuro-ophthalmic disorders. *Diagnostics.* 2023;13:160.
30. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020:1–12.
31. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis.* 2015;115:211–252.
32. Cui X, Goel V, Kingsbury B. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Trans Audio Speech Lang Proc.* 2015;23:1469–1477.
33. Asaoka R, Tanito M, Shibata N, et al. Validation of a deep learning model to screen for glaucoma using images from different fundus cameras and data augmentation. *Ophthalmol Glaucoma.* 2019;2:224–231.
34. Jiang Y, Pan J, Yuan M, et al. Segmentation of laser marks of diabetic retinopathy in the fundus photographs using lightweight U-Net. *J Diabetes Res.* 2021;2021:1–10.
35. Jumb V, Sohani M, Shrivas A. Color image segmentation using K-means clustering and Otsu's adaptive thresholding. *Int J Innovative Technol Exploring Eng.* 2014;3(9):72–76.
36. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, 2021:10012–10022.
37. Brady M, Gerhardt LA, Davidson HF. *Robotics and artificial intelligence, Vol. 11*. Springer Science & Business Media; 2012.