Neural Computing and Applications (2024) 36:23149–23168
https://doi.org/10.1007/s00521-024-10174-6

ORIGINAL ARTICLE

VISTA: vision improvement via split and reconstruct deep neural network for fundus image quality assessment

Saif Khalid[1,2] · Saddam Abdulwahab[1] · Oscar Agustín Stanchi[3,4] · Facundo Manuel Quiroga[3,5] · Franco Ronchetti[3,5] · Domenec Puig[1] · Hatem A. Rashwan[1]

Received: 18 February 2024 / Accepted: 1 July 2024 / Published online: 9 October 2024

Abstract

Widespread eye conditions such as cataracts, diabetic retinopathy, and glaucoma impact people worldwide. Ophthalmology uses fundus photography for diagnosing these retinal disorders, but fundus images are prone to image quality challenges. Accurate diagnosis hinges on high-quality fundus images. Therefore, there is a need for image quality assessment methods to evaluate fundus images before diagnosis. Consequently, this paper introduces a deep learning model tailored for fundus images that supports large images. Our division method centres on preserving the original image's high-resolution features while maintaining low computing and high accuracy. The proposed approach encompasses two fundamental components: an autoencoder model for input image reconstruction and image classification to classify the image quality based on the latent features extracted by the autoencoder, all performed at the original image size, without alteration, before reassembly for decoding networks. Through post hoc interpretability methods, we verified that our model focuses on key elements of fundus image quality. Additionally, an intrinsic interpretability module has been designed into the network that allows decomposing class scores into underlying concepts quality such as brightness or presence of anatomical structures. Experimental results in our model with EyeQ, a fundus image dataset with three categories (Good, Usable, and Rejected) demonstrate that our approach produces competitive outcomes compared to other deep learning-based methods with an overall accuracy of 0.9066, a precision of 0.8843, a recall of 0.8905, and an impressive $F1$-score of 0.8868. The code is publicly available at https://github.com/saifalkhaldiurv/VISTA_-Image-Quality-Assessment.

Keywords Autoencoder network · Explainability · Fundus image · Gradability · Interpretability · Quality assessment · Retinal image

✉ Saif Khalid
saif.khalid@qu.edu.iq;
Saifkhalidmusluh.al-khalidy@urv.cat

Saddam Abdulwahab
saddam.abdulwahab@urv.cat

Oscar Agustín Stanchi
ostanchi@lidi.info.unlp.edu.ar

Facundo Manuel Quiroga
fquiroga@lidi.info.unlp.edu.ar

Franco Ronchetti
fronchetti@lidi.info.unlp.edu.ar

Domenec Puig
domenec.puig@urv.cat

Hatem A. Rashwan
hatem.abdellatif@urv.cat

[1] Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Cataluña, Spain

[2] University of Al-Qadisiyah, Al Diwaniyah, Al-Qadisiyah 58002, Iraq

[3] Facultad de Informática, Instituto de Investigación en Informática LIDI, Universidad Nacional de La Plata (UNLP), 1900 La Plata, Buenos Aires, Argentina

[4] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), 1900 La Plata, Buenos Aires, Argentina

[5] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA), 1900 La Plata, Buenos Aires, Argentina

1 Introduction

Diagnosing conditions like diabetic retinopathy (DR) [1] and other retinal diseases relies heavily on fundus imaging [2]. However, inadequate fundus images may contain artefacts or inconsistencies, leading to misinterpretations or incorrect diagnoses when processed automatically [3]. Assessing image quality beforehand reduces the risk of inaccurate outcomes, ensuring efficient use of computational resources and time by focusing on analysing high-quality images with automated diagnostic tools.

High-resolution retinal scans play a pivotal role in clearly presenting crucial components such as the optic disc, arteries, and lesions, enabling ophthalmologists and autonomic systems to make informed clinical judgments. Early diagnosis is imperative to treat patients before the disease progresses beyond a manageable threshold. Consequently, acquiring good, clear, high-resolution images early is essential. Low-quality images can confuse reviews due to the excessive brightness and colour range in retina images, reducing the contrast between target areas (e.g. vessels and lesions) and the background. This, in turn, can befuddle ophthalmologists during fundus assessments and pose significant challenges for computer-aided retinal image analysis systems. Approximately 25% of retinal images are deemed unfit for diagnosis due to their low quality [4]. As a result, there is a growing need for Retinal Image Quality Assessment (Retinal-IQA). In clinical practice, optometrists typically perform this assessment manually, which is time-consuming and highly dependent on the operator's experience. Hence, automated Retinal-IQA is essential to streamline the process of retinal image acquisition. The primary objective of IQA in this context is to categorize images into three classes (based on quality metrics [5]): *Good*, *Usable*, and *Rejected*. Various techniques have been employed for Retinal-IQA over the years.

Early approaches involved template matching [6] and intensity-based histogram analysis using high-quality retinal images as templates. Later methods, such as those proposed by Dias et al. [7] and Wang et al. [8], focused on training classifiers for retinal image quality grading, considering various distortions. However, these techniques have limitations in encoding quality measurements effectively. Contemporary convolutional neural networks (CNNs), such as Inception v3 [9], have revolutionized Retinal-IQA. Colour space fusion networks like MCF-Net [5] integrate information from different colour spaces for Retinal-IQA. Multitask frameworks, such as MFIQA [10], enhance Retinal-IQA by incorporating additional tasks. However, it is important to note that most methods utilize low-resolution images (e.g. 512×512) with the developed models to reduce model complexity. The use of low-resolution images severely impacts the visual features of the fundus images, leading to inadequate assessments and potentially hindering overall performance.

In this paper, we aim to tackle the challenges of large image sizes and unclear visual details. To address these issues, we propose a comprehensive deep learning framework that consists of two sequential networks. A CNN network is employed to extract local and global features from the retinal images. Subsequently, these features are fed into a CNN-based classifier. The primary function of the encoder in our framework is to extract crucial features about the quality of the retinal images. These extracted features are then utilized as input for the classifier, which is responsible for categorizing the quality of the retinal images.

Current deep learning-based fundus image grading methods often lack transparency, leading to a trust deficit among ophthalmologists. Explainable deep learning models assist in understanding the decision-making process and identifying abnormal regions in fundus images, which is crucial in the medical field [11]. Despite advances in retinal image quality assessment, interpretability tools for these networks remain limited [12, 13]. Our paper proposes integrating various off-the-shelf interpretability techniques to generate interpretable visual feedback, enabling a comprehensive understanding of our model's grading process and facilitating image quality improvement. Our post hoc and intrinsic dual categorization uncovers deep learning models' decision-making in fundus image quality assessment, aiding in revealing the relationships between model inputs, internal representations, and gradability classifications, addressing the critical need for transparency and comprehension in medical image analysis. Figure 1 illustrates the structure of our proposed method termed "VISTA" for image gradability classification. The main contributions of this work can be summarized as follows:

- Introducing a novel "data streaming" approach that allows input images to be processed without the need for resizing, thus preserving the original large image size of 1280×1280. This strategy reduces the training cost, enhances model efficiency, and conserves computer resources.
- Proposing a multi-layer deep network that can extract both local and global features from images. This design facilitates the capture of intricate details, thereby improving image classification performance.
- Integrating the two networks into a unified pipeline, which combines the extracted local and global features. These merged features are subsequently utilized by the decoder to reconstruct the input image, thereby supporting the learning process of our model.

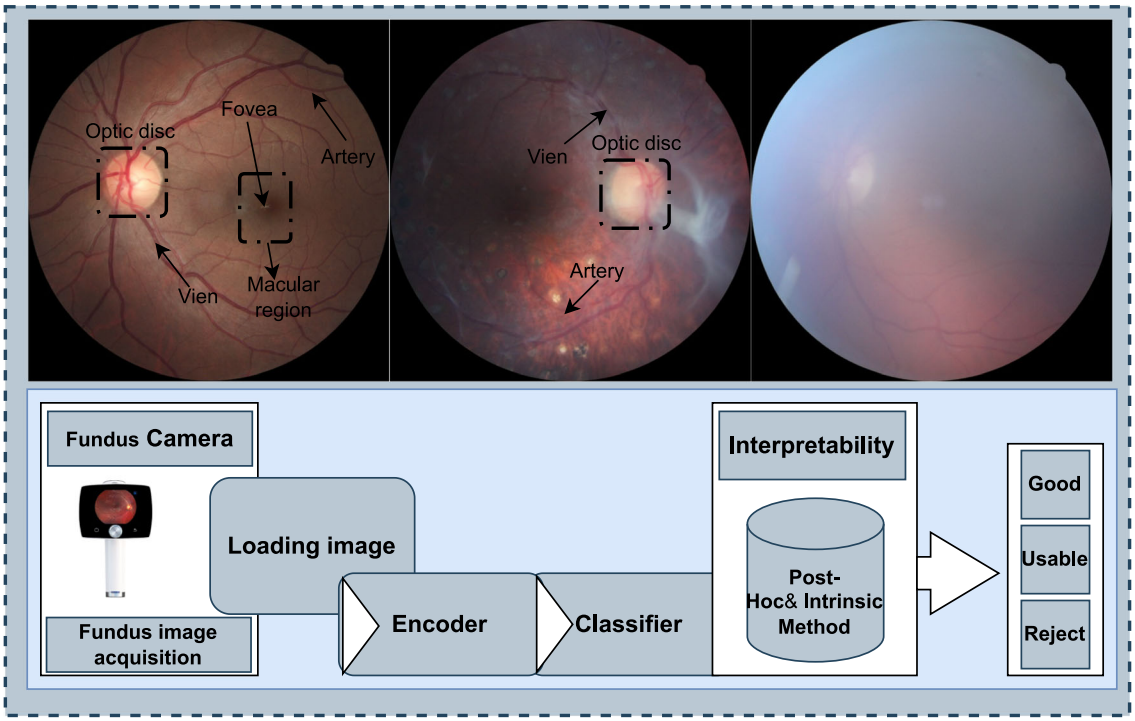


Fig. 1 An overview of the VISTA model. Retinography generates a fundus image, which our model processes. The outputs gradability results and interpretability feedback to ophthalmologists

- Introducing a convolutional neural network (CNN) classifier that effectively utilizes the latent features learned by the encoder network to classify fundus images as either gradable or ungradable.
- Implementing a composite error approach that combines a mean squared error (MSE) loss with standard categorical cross-entropy (CE) during the training of the autoencoder model. Additionally, for the classification task, we employ the CE loss function in conjunction with the CNN classifier to categorize fundus images based on their gradability status.
- Integrating between post hoc and intrinsic interpretability methods for the developed model to understand and validate the classification process of the developed deep learning model for ensuring its transparency and reliability in clinical settings.

The structure of this paper is as follows: Sect. 2 discusses the related work. Section 3 outlines our proposed methodology to categorize fundus images as gradable or ungradable using both networks. Section 4 offers an in-depth methodological explanation for the interpretability of our model. Section 5 presents experimental results and performance evaluations. Section 6 presents the results and discussion. Finally, in Sect. 7, we conclude our work and suggest potential avenues for future research.

2 Related work

As mentioned in the previous section, deep learning-based fundus IQA methods have emerged as powerful tools for automatically assigning quality grades to fundus images. Leveraging intricate neural network architectures such as U-Net, ResNet, and EfficientNet, these methods are capable of capturing subtle patterns and features, thereby enabling a nuanced evaluation of image quality. These advancements are crucial for addressing the challenges highlighted in the introduction, particularly the need for accurate and efficient image quality assessment to support early diagnosis and treatment of retinal diseases. For a comprehensive overview of recent developments in fundus IQA methods, refer to Table 1.

Several studies for IQA of fundus images, including those proposed in [5] and [14], adopt a binary classification system, categorizing image quality as either *Good* or *Rejected*. However, this approach has inherent subjectivity, reliant on specific diseases and the discretion of ophthalmologists. Moreover, the binary labels fail to represent the continuum of image quality, potentially leading to classification errors. A numerical estimation of image quality could bolster diagnostic confidence. While some studies focus on specific attributes like focus and contrast, they often overlook critical elements such as macular and optic disc visibility. Additionally, benchmarking algorithm

Table 1 Comparisons of the proposed methods in state-of-the-art on Eye-Quality. (Inter) refers to Interpretability, ($F1$) refers to $F1$-score

#	No. of class	Image size	Model and references	$F1$	Contribution	Inter
1	3	224 × 224	MCF-Net [5]	0.8551	Re-annotate an EyeQ dataset and analyse the influences on Retinal-IQA of different colour spaces (DenseNet121v 3v 1)	NO
2	2	224 × 224	NBIQA [21]	0.7441	Hand-crafted feature-based methods simultaneously consider the features from both the spatial and transform domains	NO
3	2	224 × 224	BRISQU [22]	0.7112	Hand-crafted feature-based methods adopt local normalization brightness coefficients for image quality	NO
4	2	224 × 224	TS-CNN [23]	0.7481	Natural-IQA includes an image stream focusing on grey information structure stream focusing on image gradient details	NO
5	2	224 × 224	HVS [8]	0.6991	Hand-crafted methods apply HVS features, multichannel sensation, noticeable blur, and contrast sensitivity measure	NO
6	3	512 × 512	MR-CNN [16]	0.8694	CNN consists of 4 pre-trained models (Inception-V3, ResNet-151, DenseNet-121, Xception) to derive the optimized features	NO
7	3	224 × 224	DenseNet121 [5]	0.8551	Dense blocks and transition layers exist between two adjacent dense blocks along with a global average pooling layer	NO
8	3	224 × 224	ResNet-18 [5]	0.808	Residual blocks, skip connections, basic building blocks, pooling, and fully connected layers are included	NO
9	3	224 × 224	Single-branch SalStructIQA [14]	0.8662	A single CNN branch first fuses the retinal image and salient structures	Yes
10	3	224 × 224	Dual-branch SalStructIQA [14]	0.8723	Two parallel CNN branches are used for deep feature learning	NO
11	2	224 × 224	CNN combined [19]	0.878	A CNN model is used to assess the quality and combine deep and generic texture features, using a RandomForest classifier	NO
12	2 &3	480 × 480	FGR-Net(Autoencoder-vgg-16) [15]	0.8782	A deep autoencoder reconstructs the input image to extract the visual characteristics based on self-supervised learning	Yes
13	2 &3	224 × 224	MFQ-Net [24]	0.8564	A deep learning-based model in a smartphone consists of two main blocks: PFE and IC block	NO
14	3	224 × 224	SG-Net [25]	0.8676	The model consists of a vessel segmentation module (VSM), an optic disc segmentation (ODSM), a quality assessment module (QAM), and uses U-Net	NO
15	3	224 × 224	blood vessels(ResNet-50-p) [26]	0.7967	An end-to-end learning-based method is used for segmenting the blood vessels of the input image by U-Net	Yes
16	3	224 × 224	UW-OCTA [27]	–	A fully automated convolutional neural network-based method (EfficientNet-B2) is presented	NO
17	2	224 × 224	Retinal image quality [28]	0.7572	A CNN pre-trained on non-medical images is used for extracting general image features	NO
18	2	224 × 224	MRDB-CNN [29]	0.7697	A modified residual dense block convolution neural network (MRDB-CNN) uses ResNet-34	NO
19	2	224 × 224	A siamese network [30]	0.7830	A siamese network with two weight-shared branches is used to compare the quality of two images of the same scene	NO
20	2	224 × 224	SCNN [20]	0.7568	A CNN based on a shuffle unit mixes up the extracted features	NO

performance remains challenging, with varied labelling standards hindering the evaluation of generalizability.

Furthermore, most IQA methods rely on low-resolution fundus images to streamline the developed deep learning models. Table 1 shows many works using image sizes of 224×224 [5], 480×480 [15] or maximum 512×512 [16]. However, utilizing such low-resolution images for assessment compromises critical details, impeding accurate identification and a comprehensive understanding of image quality. The resulting loss of intricate visual information distorts crucial components, potentially leading to misinterpretations of ocular irregularities and incorrect diagnostic conclusions. Additionally, the restricted resolution limits the capture of subtle textures and colour nuances, impacting the sensitivity and specificity of quality assessment. Therefore, low-resolution images present a

significant challenge to achieving precise and dependable quality assessment, undermining the effectiveness and accuracy of clinical evaluations. Incorporating high-resolution images can enhance the performance of deep learning models. This can be achieved through a "data streaming" approach via extracting local features in the input image and then merging the features to get global image representation, facilitating the processing of input images with larger dimensions [17, 18].

In addition, the majority of deep learning-based methods for image quality assessment depend on CNN for classification [19]. However, their efficacy diminishes when dealing with fundus images from different cameras, as CNNs merely replicate the training data. Expanding the convolutional layers in the encoder network can result in the loss of information, leading to misclassifications by the classifier layer. An autoencoder network compresses input features before quality classification to tackle this issue, as described in [15]. The autoencoder prioritizes essential image characteristics and learns significant properties from the data. Furthermore, the encoder functions as a feature extractor, transmitting latent features to the classifier network, thereby encoding all crucial visual attributes necessary for accurate IQA.

Deep learning-based methods for image quality assessment have often operated opaquely, lacking interpretability in their decision-making, which limits clinicians' trust [20]. The opacity of many models has contributed to scepticism due to the absence of transparency. Therefore, integrating interpretability tools into image quality assessment is vital to illuminate the intricate features influencing decisions. Interpretability enhances trust by elucidating the decision-making process, allowing clinicians valuable insights into factors contributing to quality assignments [15]. Incorporating both post hoc and intrinsic interpretability methods is essential for enabling explainability in the developed model, providing a comprehensive understanding of how the model arrives at specific quality assessments. This approach addresses the historical limitation of opacity and improves the acceptance and reliability of deep learning-based image quality assessment methods in medical diagnostics.<Table ID="

3 Proposed deep learning model

3.1 VISTA (Split and reconstruct image for fundus image quality assessment)

This section introduces the proposed model, VISTA, for assessing the gradability of retinal images. Additionally, interpretability techniques are presented to offer visual insights into the criteria used by VISTA for gradability classification. Figure 2 illustrates our model for evaluating the gradability of fundus images. Initially, the model splits the original large-size image (1280×1280) into $n \times n$ segments without resizing for training. This approach facilitates training on large, high-resolution images, reducing memory usage and processing time. Consequently, the $n \times n$ smaller sub-images or patches will be individually processed during training using shared CNN blocks, with variations in size to accommodate the model's requirements and memory capacity. Parallel processing on CPUs or GPUs accelerates the training process. This strategy effectively manages memory demands, enabling training on high-resolution images without encountering memory limitations. The model operates as follows.

3.1.1 Encoder

The encoder component is employed to extract both local and global features from images. Local features refer to details and patterns within small image regions (e.g. $n \times n$ sub-images). Our model architecture adeptly captures local spatial patterns using filters or kernels, scanning the sub-image to learn features like edges, corners, and blobs. The third layer of the model generates a collection of feature maps that represent the image's local features for each sub-image. The local feature extraction is achieved using shared $n \times n$ convolutional blocks. Global features encompass the overall image structure and context by amalgamating all local features into one comprehensive feature map. Subsequent layers aim to extract high-level features (i.e. latent features) that signify the crucial visual attributes of the input fundus image. Table 2 shows the detailed structure of the encoder network.

3.1.2 Decoder

The decoder architecture in our model plays a pivotal role in reconstructing high-resolution images from abstracted feature representations. The decoder architecture is defined by a sequence of seven blocks. The block in the provided neural network architecture is a pivotal component responsible for upsampling, contributing to the hierarchical feature refinement in the decoding process. This block has two main elements: a transposed convolutional layer and a convolutional block. The transposed convolutional layer facilitates a 2×2 upsampling, effectively doubling the spatial dimensions of the input. This operation is crucial for recovering finer spatial details lost during earlier downsampling stages. The convolutional block module in the provided neural network architecture is a fundamental feature extraction and processing building block. Comprising convolutional layers, batch normalization, and rectified linear unit (ReLU) activation, this block captures

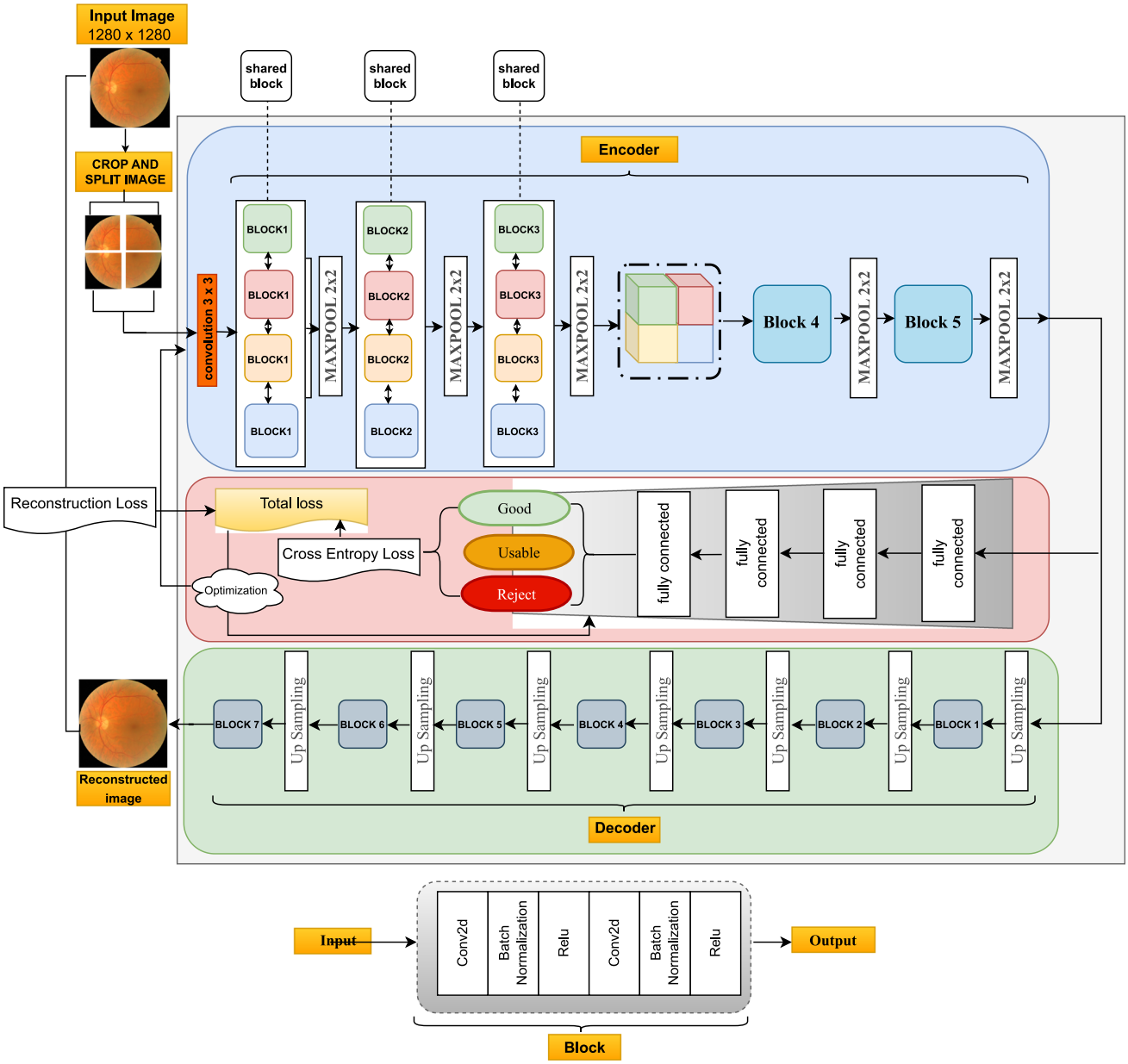


Fig. 2 General overview of the proposed model in train and test stage

and enhances intricate features within the input data. Employing transposed convolutional layers and convolutional blocks to facilitate upsampling and feature refinement. Its role is crucial in achieving the high-fidelity reconstruction of the original fundus images in the broader context of medical image analysis. For a comprehensive understanding of the structural configuration and the specific details of each layer within the decoder block, refer to the elaborated information provided in Table 3. This table offers a detailed breakdown of the kernel sizes, layer types, and other pertinent architectural aspects, providing valuable insights into the decoder.

3.1.3 Classifier

The resulting feature map (i.e. latent feature) from the encoder network, sized at $15 \times 15 \times 512$, serves as input for a classifier network responsible for categorizing retinal fundus image quality into distinct gradability categories. The classifier network consists of four fully connected

Table 2 The detailed structure for input and output of the encoder network

Layer	Scales	Input size	Output size
Input Image (1280 × 1280)			
Block 1	Scale 1	3 × 640 × 640	64 × 320 × 320
	Scale 2	3 × 640 × 640	64 × 320 × 320
	Scale 3	3 × 640 × 640	64 × 320 × 320
	Scale 4	3 × 640 × 640	64 × 320 × 320
Block 2	Scale 1	64 × 320 × 320	128 × 160 × 160
	Scale 2	64 × 320 × 320	128 × 160 × 160
	Scale 3	64 × 320 × 320	128 × 160 × 160
	Scale 4	64 × 320 × 320	128 × 160 × 160
Block 3	Scale 1	128 × 160 × 160	256 × 80 × 80
	Scale 2	128 × 160 × 160	256 × 80 × 80
	Scale 3	128 × 160 × 160	256 × 80 × 80
	Scale 4	128 × 160 × 160	256 × 80 × 80
Block 4	Scale 1	256 × 80 × 80	512 × 40 × 40
Block 5	Scale 1	512 × 40 × 40	512 × 10 × 10

Table 3 The detailed structure for input and output of the decoder network

Conv	Input size	Output size
Input Image (1280 × 1280)		
Upsample-1	512 × 10 × 10	512 × 20 × 20
Block 1	512 × 20 × 20	512 × 20 × 20
Upsample-2	512 × 20 × 20	512 × 40 × 40
Block 2	512 × 40 × 40	512 × 40 × 40
Upsample-3	512 × 40 × 40	256 × 80 × 80
Block 3	256 × 80 × 80	256 × 80 × 80
Upsample-4	256 × 80 × 80	128 × 160 × 160
Block 4	128 × 160 × 160	128 × 160 × 160
Upsample-5	128 × 160 × 160	64 × 320 × 320
Block 5	64 × 320 × 320	64 × 320 × 320
Upsample-6	64 × 320 × 320	32 × 640 × 640
Block 6	32 × 640 × 640	32 × 640 × 640
Upsample-7	32 × 640 × 640	32 × 1280 × 1280
Block 7	32 × 1280 × 1280	3 × 1280 × 1280

Table 4 The detailed structure of the classifier network. The output of classifier4 depends on the number of classes

Layers	Layer type	Input features	Output features	Bias
Classifier1	Linear	512	256	True
Classifier2	Linear	256	128	True
Classifier3	Linear	128	64	True
Classifier4	Linear	64	No. of classes	True

layers, followed by a ReLU activation layer. Table 4 has a detailed outline of the classifier network's architecture.

3.2 Training

This study employs two loss functions to enhance the performance of the developed model. The first function is dedicated to the classification task (i.e. the classification loss). In contrast, the second function is utilized for self-supervision training (i.e. the reconstruction loss).

Initially, we assessed the model's performance using different reconstruction loss functions (L_{rec}). We then chose the most effective model to compare input images with their corresponding reconstructed counterparts within self-supervised learning. We explored standard loss functions to underscore the importance of the autoencoder architecture in enabling the network to identify pertinent patterns related to image quality characteristics, irrespective of the specific loss function chosen.

The first reconstruction loss was evaluated for the training of the autoencoder network as an MSE, L_{MSE}, computed from the features extracted from the input image I and the reconstructed image $\hat{I}$. MSE is a widely used loss function for regression tasks, measuring the mean of the squared differences between actual and predicted values. It is defined as follows:

$$L_{\text{MSE}}(I,\hat{I}) = \frac{1}{n}\sum_{i=1}^{n}(I_i - \hat{I}_i)^2, \tag{1}$$

where $I_{(i)}$ is the input image of pixel i, $\hat{I}_{(i)}$ is the reconstructed image and the n is the numbers of pixels in an image.

The second tested reconstruction loss function, L_{VGG}, was the Perceptual Loss, also known as VGG Loss. It measures the similarity between the features of the input image and the reconstructed images obtained from a pre-trained convolutional neural network, typically a VGG network. This loss function assesses the ability of the reconstructed image to capture the crucial features of the original input image, as identified by the pre-trained neural network.

$$L_{\text{VGG}}(I,\hat{I}) = \frac{1}{N_{\text{m}}}\sum_{i=1}^{N_m} ||\phi_i(I) - \phi_i(\hat{I})||_2^2, \tag{2}$$

where ϕ_i is the feature map extracted by the VGG network at the ith layer, and N_{m} is the total number of feature maps. The loss function calculates the mean squared difference between the feature maps of the input and reconstructed images using the L2-norm. $\|.\|_2$ represents the L2-norm, and $\|.\|^2$ denotes the square of the L2-norm, applied to the

difference between the corresponding feature maps at each layer.

The third reconstruction loss function, L_{SSIM}, is derived from the structural similarity index Measure (SSIM), a technique used for assessing the perceived quality of digital images. SSIM quantifies the similarity between two images, serving as a comprehensive metric for evaluating the quality of reconstructed images compared to their corresponding input images.

In contrast to L1 and L2 loss functions emphasizing pixel-level disparities, the SSIM metric assesses image likeness based on essential parameters: Luminance, contrast, and structure. Thus, SSIM is a well-established measure for quantifying the distinctions between two images.

SSIM can be mathematically defined as:

$$L_{\text{SSIM}}(I,\hat{I}) = \frac{(2\mu_{\hat{I}}\mu_I + c_1)(2\sigma_{\hat{I}I} + c_2)}{(\mu_{\hat{I}}^2 + \mu_I^2 + c_1)(\sigma_{\hat{I}}^2 + \sigma_I^2 + c_2)}, \tag{3}$$

where $\mu_{\hat{I}}$ is the mean of $\hat{I}$, $\sigma_{\hat{I}}$ is the standard deviation of $\hat{I}$, μ_I is the mean of I, σ_I is the standard deviation of I, $\sigma_{\hat{I}I}$ is the covariance of $\hat{I}$, $c_1 = 0.01^2$ and $c_2 = 0.03^2$, as described in [31]. The choice of these constants is based on extensive experimentation and evaluation across various image processing applications, where they have demonstrated effectiveness in balancing the trade-off between sensitivity to structural information and numerical stability.

The fourth reconstruction loss function, denoted as L_{PSNR}, is based on the peak signal-to-noise ratio (PSNR). PSNR is a widely employed loss function for image reconstruction, serving as a metric to quantify the dissimilarity between the input image and the reconstructed image in terms of the peak signal-to-noise ratio. PSNR is commonly used to assess the quality of the reconstructed image by comparing it to the original input image.

$$L_{\text{PSNR}}(I,\hat{I}) = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right), \tag{4}$$

where MAX_I is the maximum possible pixel value (i.e. 255 for an 8-bit image), and MSE is the mean squared error between the input image I and the reconstructed image $\hat{I}$.

The PSNR measures the ratio of the maximum possible power of a signal to the power of corrupting noise that affects the fidelity of its representation. It is often used to evaluate the quality of reconstructed images.

The fifth reconstruction loss function, denoted as L_{TL} and known as the Tversky loss function, is typically utilized in segmentation tasks, where the primary goal is to predict a binary mask indicating the presence or absence of an object in each pixel of an image. However, it is possible to adapt this loss function for image reconstruction tasks. Here is one approach to defining the Tversky loss function for image reconstruction:

$$L_{\text{TL}_{\alpha,\beta}}(I,\hat{I}) = \frac{\sum_i I_i\hat{I}_i}{\sum_i I_i\hat{I}_i + \alpha\sum_i I_i(1-\hat{I}_i) + \beta\sum_i(1-I_i)\hat{I}_i}, \tag{5}$$

where x and $\hat{x}$ are the input and reconstructed images, respectively, and N is the total number of pixels in the image (each summation in the loss function is taken over all N pixels in the image). In this formula, the Tversky loss function measures the similarity between the ground truth and predicted images, with larger values indicating more significant similarity. The weight factors α and β can be used to balance the importance of the false positive and false negative rates in the loss function, similar to the segmentation case.

The second loss function for the classification task, used for the quality labelling task, is denoted as L_{CE} and is based on CE. This loss function depends on the predicted class from the classifier $\hat{y}$ and the corresponding target value y. The CE loss, L_{CE}, is defined as follows:

$$L_{\text{CE}}(\hat{y_i}, y_i) = -\sum_{i=1}^{n} y_i \cdot \log(\hat{y_i}), \tag{6}$$

where $\hat{y_i}$ represents the ith scalar value in the classifier output, y_i is the corresponding target value, and the output size indicates the number of scalar values in the model output. This loss is an excellent measure of how distinguishable two discrete probability distributions are from each other. In this context, y_i represents the probability of event i occurring, and the sum of all y_i equals 1, signifying that exactly one event may occur. The minus sign ensures that the loss decreases as the distributions approach each other.

The final objective loss function, L, to optimize the VISTA model, including the autoencoder and classifier networks, is a combination between the reconstruction loss L_{rec} (i.e. one of the aforementioned loss functions yields the best performance) and the classification loss function L_{c}, as:

$$L = \alpha L_{\text{rec}}(I,\hat{I}) + (1-\alpha)L_c(\hat{y_i}, y_i), \tag{7}$$

where α is a weight factor set to 0.5 in this work.

4 Methodological overview of interpretability for the VISTA model

Our proposed approach encompasses a dual categorization of interpretability methods, a pivotal aspect in elucidating the decision-making processes of deep learning models in

fundus image quality assessment. These methods fall into two overarching groups: post hoc and intrinsic.

The integration of both post hoc and intrinsic methods in our study allows for a comprehensive exploration of the interpretability landscape in the context of fundus image quality assessment. This dual-method strategy equips us to elucidate the intricate relationships between model inputs, internal representations, and the ultimate gradability classifications, addressing the imperative need for transparency and comprehension in medical image analysis.

For reasons of brevity, in this section, we only focus on the *Good* and *Rejected* classes of the dataset, which represent 81% of its samples.

4.1 Post hoc approaches

Attribution methods are a subtype of post hoc methods that gauge the significance of individual components by introducing modifications to the input or internal elements and subsequently observing the resulting impact on the model's performance. Attribution methods provide insights into which features the model considers decisive for its gradability predictions, thereby enhancing transparency.

In this work, we leverage three prominent attribution methods: Gradient, Grad-CAM, Occlusion, and RISE.

The Gradient or Saliency Maps method provides insights into feature importance by measuring the sensitivity of the model's output concerning changes in individual pixel values [32]. This approach involves calculating the gradient of the class score concerning the input fundus image. Green values in the resulting visualizations indicate that increasing a pixel's brightness contributes to a higher model score. In turn, Grad-CAM is another gradient-based technique that leverages the gradients of the target class by streaming them into the final convolutional layer. This process results in a coarse localization map highlighting crucial regions in the fundus image for predicting the target class [33]. The output of Grad-CAM is usually interpreted as the importance of each region for a specific class (i.e. in or case two classes), regardless of sign. The method visually highlights crucial areas for the model's decision-making process.

In perturbation-based approaches, the Occlusion method quantifies the impact of systematically occluding various segments of the input image on the model's class scores [32]. Green values signify regions where occlusion increases the class score, while red values indicate the opposite. However, Occlusion can be very compute-intensive for larger images and suffers from biases in choosing the occluded area's shape and size. Therefore, Occlusion's result can be improved upon by RISE (Randomized Input Sampling for Explanation), a perturbation-based approach that employs a Monte Carlo integral approximation algorithm to generate pixel importance maps [34]. It samples random occlusion masks, considers spatial structure, and produces saliency maps by combining sub-sampled shows.

4.2 Intrinsic approaches

Intrinsic methods, on the other hand, endeavour to enhance the interpretability of internal representations by incorporating techniques that are inherently part of the VISTA model's architecture. They allow training of traditional black box models in such a way that their internal representations are more interpretable. Unlike attribution methods, intrinsic approaches focus on refining the interpretability within the model itself, thereby elevating fidelity, clarity, and parsimony in the attribution of importance to specific features.

For intrinsic interpretation, our model is trained using Concept Whitening (CW) [35], aligning model representation with key concepts, thus enhancing interpretability. To apply CW to the VISTA architecture, the layer was placed in Block 4 after the last convolution operation.

CW shapes the latent space of a model by compelling it to learn the representation of key concepts using concept vectors [35]. Similarly to a whitening transformation, CW aligns the axes of the latent space with known auxiliary concepts of interest. Concepts are user-defined and facilitate the separability of the latent space, enhancing the interpretability of the model's internal representations. Using CW in a network requires minimal modifications to the architecture itself.

During inference, the activation pattern CW layer quantifies those concepts as an auxiliary explanation.

5 Experiments

5.1 Eye-quality dataset

There are several publicly available Retinal-IQA datasets with manual quality annotations, such as HRF [7], DRIMDB [13], and DR2 [11]. However, these datasets have various drawbacks. First, more than one dataset is based on binary labels, i.e. *Good* and *Rejected* without any intermediate level. However, several images fall between these two categories and are essential for classification. For example, some fundus images with poor quality, containing a few artefacts, or are slightly blurred, are still gradable, so they should not be labelled as *Rejected*. Still, they may mislead automated medical analysis methods, so they cannot be labelled as *Good*. Second, retinal images of the existing Retinal-IQA datasets are often captured by the same camera, which cannot be used to evaluate the

robustness of Retinal-IQA methods against various imaging modalities. Third, existing datasets are limited in size; large-scale quality grade datasets are lacking for developing deep learning methods.

In our paper, we use the Eye-Quality (EyeQ) dataset [5] from the EyePACS dataset, which has three classes (*Good*, *Usable*, and *Rejected*) and a large image of 1280×1280. EyeQ contains 26,000 samples captured by different models and types of cameras under various imaging conditions. It is divided into a training set of 12,417 images, a test set of 13,430 images, and a validation set of 153 images. The dataset utilizes a three-level quality grading system by considering four common quality indicators: blurring, uneven illumination, low contrast, and artefacts. The three quality grades are defined as follows: (1) A retinal image is graded as *Good* if it exhibits no low-quality factors and all retinopathy characteristics are visible. (2) Alternatively, it is graded as *Usable* if it displays some slight low-quality indicators, such as low contrast, blurriness or artefacts, which may affect automated medical analysis methods. However, the main structures (such as the disc and macula regions) and lesions must still be discernible by ophthalmologists. For cases where uneven illumination is present, a retinal image is considered *Usable* if the readable region of the fundus image is more significant than 80% of the total image area. (3) A retinal image is graded as *Rejected* if it exhibits serious quality issues that render it unsuitable for providing a complete and reliable diagnosis, even by ophthalmologists. A fundus image with no visible disc or macula region is also classified as *Rejected*.

5.2 Experimental setup

In our experiments, we employed a dedicated setup for our data processing pipeline.[1]

5.2.1 Data augmentation

Deep neural networks depend on the training data available to achieve optimal performance. Additionally, the dataset requires increasing training data for a stable model because unbalanced data will lead to overfitting the EyeQ dataset. After augmentation, our dataset contains 22,996 in retinal images for training. Figure 3 shows some examples of the augmentation applied to each input image. The transformations employed Crop, Flip, Rotate, and ColorJitter.

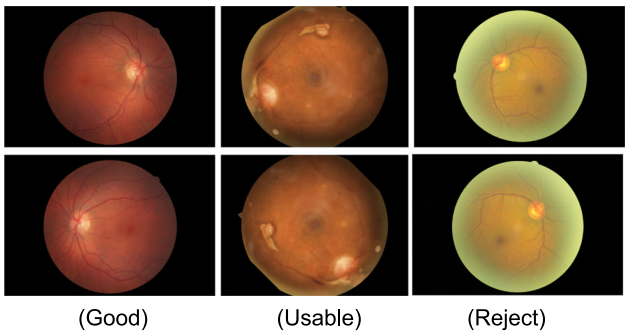


Fig. 3 Examples of transformations applied to an input image of each class of the EyeQ dataset

Figure 4 shows the class distribution before and after augmentation.

5.2.2 Hyperparameters

We used the stochastic gradient descent (SGD) with $\gamma = 0.1$ and an initial learning rate of 0.001. A batch size of 2 and 50 epochs yielded the best combination. For a more detailed overview of the hyperparameters setting of our proposed network architecture, please refer to Table 5.

5.3 Evaluation measures

We used four metrics to measure the resulting performance: accuracy, sensitivity, specificity, and $F1$-score. In medical diagnosis, the most important measure of the model's performance is the sensitivity measure, which refers to our model's ability to correctly identify high-quality images, i.e. measuring the percentage of true positives. In contrast, specificity measures the model's ability to correctly identify low-quality images.

Table 6 compares class-specific evaluation metrics for the VISTA model across the three classes (*Good*, *Usable*, and *Rejected*). Notably, the VISTA model demonstrates exceptional precision, recall, and $F1$-score in the *Good* and *Rejected* classes, indicating robust performance. The *Usable* class, presenting a challenge due to its intermediary nature between *Good* and *Rejected*, still achieves acceptable results. In summary, the VISTA model attains an overall Accuracy of approximately 0.91, with a precision of 0.88, recall of 0.89, and an impressive $F1$-score of 0.89.

5.4 Ablation study

Initially, we assessed the performance of various models using distinct backbones for the autoencoder network. Specifically, we implemented Mobilenetv2 [36], DenseNet169 [37], ResNet-50 [38], ResNeXt101 [39], CoAtNets [40], Inception-v4 [41], and VGG-16 [42]. Each network

[1] The experimental environment is a computer with an Intel Xeon E5-2667 CPU, 64 GB of RAM, and NVIDIA 1080 Ti GPU running Ubuntu 18.04. We use the PyTorch library to implement our architecture.

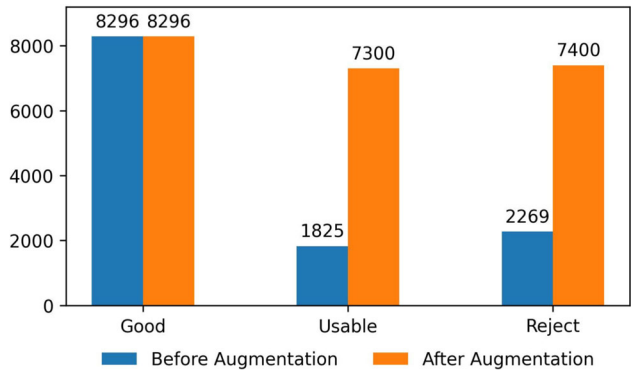


Fig. 4 Class distribution before and after data augmentation over EyeQ dataset

Table 5 Hyperparameters setting of our proposed network architecture

Hyperparameter	Value
Objective function	MSE
Optimizer	SGD
Gamma	0.1
Momentum	0.9
Learning rate	0.001
Batch dize	2
Epochs	50

Table 6 Comparison between the evaluation metrics for classification model for each class of VISTA model

Class label	Precision	Recall	*F*1-score
Good	0.95829	0.93561	0.94681
Usable	0.76813	0.83434	0.79987
Rejected	0.92654	0.90161	0.91391
Macro avg	0.88432	0.89052	0.88686

Table 7 Evaluation of the VISTA model based on different backbones for the autoencoder network on EyeQ

Method	Accuracy	Precision	Recall	*F*1-score
VISTA-Mobile net v2	0.8786	0.8584	0.8324	0.8441
VISTA-Densenet169	0.8439	0.8135	0.7742	0.7905
VISTA-Resnet-50	0.8942	0.8766	0.8776	0.8754
VISTA-ResNeXt101	0.8882	0.8723	0.8667	0.8667
VISTA-CoAtNet-1	0.8742	0.8536	0.8764	0.8621
VISTA-Inception-v4	0.8830	0.8486	0.8544	0.8479
VISTA-VGG-16	***0.9066***	***0.8843***	***0.8905***	***0.8868***

underwent training from the ground up, and the quantitative outcomes are detailed in Table 7. Notably, VGG-16 exhibited superior results among all the tested backbones. We believe that VGG-16, distinguished by its 16 convolutional layers compared to other tested networks like ResNet-50, possesses a superior capacity to discern intricate features and nuanced patterns within fundus images. Given the complex structures and subtle variations inherent in fundus images, VGG-16's architecture enables it to capture these details effectively. Additionally, our extensive experimentation and evaluation on our dataset have consistently revealed VGG-16's superiority over other backbone networks in terms of accuracy, precision, and recall for fundus image quality assessment tasks. While the precise reasons for its outstanding performance may vary depending on specific dataset characteristics, these empirical findings reaffirm the efficacy of VGG-16 as a backbone network within our research domain. Consequently, for our model (VISTA), we opted for VGG-16 as the backbone, considering its optimal performance across the four metrics outlined in Table 6.

In the next phase of our ablation study, we focused on identifying the optimal reconstruction loss function, denoted as L_{rec}, for the VISTA model. To accomplish this, we trained the VISTA model with a VGG-16 backbone, employing six different loss functions for reconstruction. These included the MSE loss, along with others such as Tversky, SSIM, VGG Perceptual, PSNR, and a composite loss function named Sum Loss. Our experiments, illustrated in Fig. 5, involved evaluating the model's performance using the EyeQ dataset. Notably, the MSE loss function outperformed the six alternative loss functions across all four evaluation metrics-Accuracy, Recall, Precision, and *F*1-score, as depicted in Table 8. Consequently, we selected MSE as the loss function to optimize the autoencoder network, aiming to minimize the error between reconstructed and input images.

5.5 Experimental setup of a post hoc interpretable VISTA model

In order to improve efficiency, we evaluate interpretability methods utilizing half-precision for faster computation and computational efficiency. We found no difference in results when using larger single or double precision.

In the execution of interpretability methods, specific parameters and considerations were carefully tailored to enhance the validity and relevance of the obtained insights. The nuances of these considerations are elucidated for each method below:

For RISE, the spectrum of values is automatically stretched to the range 0–1 for plotting (min-max normalization). In the application of Grad-CAM, the last

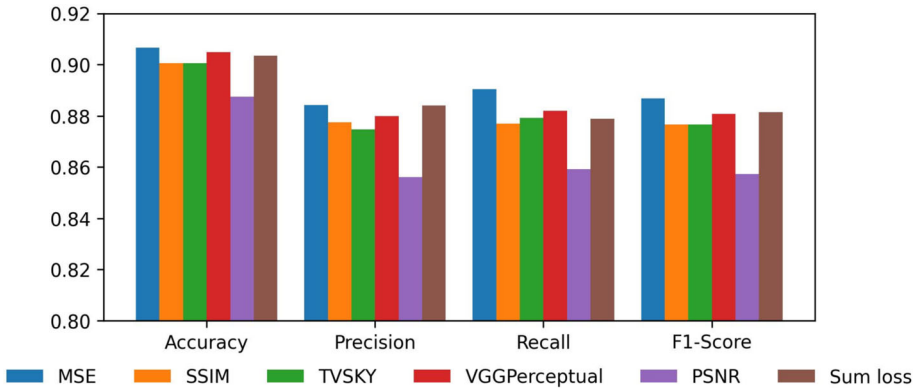


Fig. 5 Comparison of the VISTA model with six loss functions (MSE, SSIM, Tversky, VGG Perceptual, PSNR and Sum Loss) on EyeQ

Table 8 Comparison between the VISTA model with different loss functions on EyeQ

Loss function	Accuracy	Precision	Recall	*F*1-score
MSE	***0.9066***	***0.8843***	***0.8905***	***0.8868***
SSIM	0.9006	0.8775	0.8770	0.8767
Tversky	0.9006	0.8748	0.8792	0.8767
VGG perceptual	0.9048	0.8799	0.8820	0.8808
PSNR	0.8875	0.8560	0.8592	0.8573
Sum loss	0.9035	0.8841	0.8788	0.8814

convolutional was chosen as an intermediate representation for backpropagation.

Then, considerable adjustments were made to the Occlusion method to account for the larger image size by using hyperparameters adapted from a previous study [15], where the stride was set to 3x120x120, and the sliding window was configured as $3 \times 240 \times 240$.

5.6 Design of concept decomposition for an intrinsic interpretable VISTA model

To annotate the intermediate concepts, we employed LabelStudio, iterating through multiple labelling cycles and final dataset curation stages, to ensure non-contamination of the training data. The annotated concepts *illumination-bright* and *illumination-dark* were selected as they represent mutually exclusive conditions, overexposure and underexposure, respectively.

We annotated a total of 312 images, with 254 belonging to the first concept and 258 to the second one. Additionally, a concept labelled as "good" was annotated for 312 images as well to enable the model to comprehend the distinction between images of poor and good quality.

We previously experimented with using more concepts to further refine the interpretability of the method. However, we found that the CW layer is unable to learn complex decompositions of quality issues in fundus images. Figure 6 shows one failure case, where the model consistently grouped only the *illumination-bright* and *illumination-dark* concepts. We hypothesize that the underlying issue is rooted in the multiple superposed factors that contribute to the decrease of quality in fundus images. For instance, most images that suffer from low contrast also have brightness anomalies (too bright or too dark areas). Artefacts can be confounded by small areas with unusual brightness. Blurring is usually accompanied by a lack of contrast. Therefore, the concepts were deliberately selected to avoid these ambiguities.

6 Results and discussion

6.1 Comparisons with state-of-the-art models

To compare the VISTA model with modern gradability assessment methods, we compared VISTA to the state-of-the-art three class (*Good*, *Usable*, and *Rejected*) gradability assessment on the public EyeQ dataset. Table 9 summarizes the results of the VISTA with eight methods: two methods are based on handcrafted features; BRISQUE [22] and NBIQA [21], and six methods based on deep learning designed for retinal fundus image quality assessment; TS-CNN [23], HVS-based method [8], MCF-Net [5], multivariate regression CNN(MR-CNN) [16], the Double branch network, SalStructIQA [14] and multilevel quality assessment network [19]. The VISTA used VGG-16 as a backbone and MSE as a loss function. Our model demonstrates a notable enhancement across all four metrics-Accuracy, Precision, Recall, and *F*1-score-showing improvements of approximately 2, 1, 2, and 1.5%, respectively, when compared to two existing variations, SalStructIQA [14] and CNN combined [19]. The approach introduced in [14] involves segmenting two salient features before assessing fundus image quality, adding complexity to their model. Similarly, the method proposed in [19] combines deep and

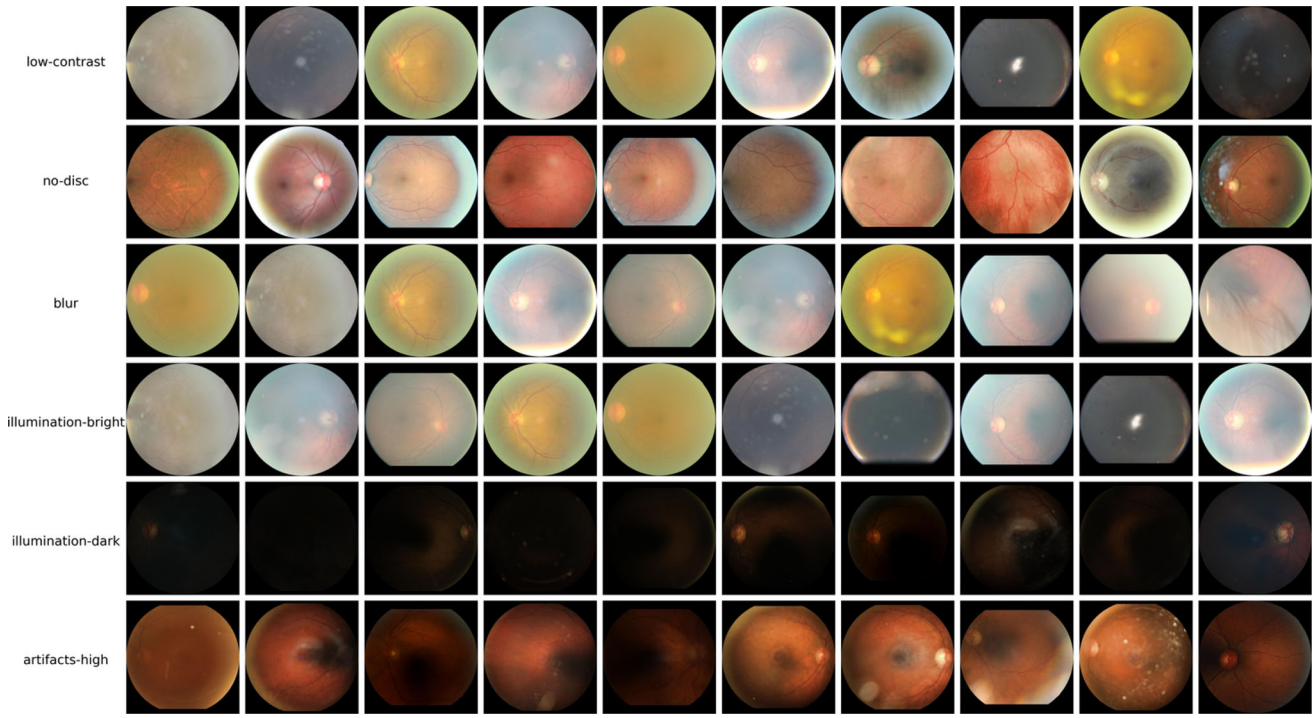


Fig. 6 Model's tendency to consistently group only the *illumination-bright* and *illumination-dark* concepts, reinforcing our intentional selection to focus exclusively on these concepts

Table 9 Comparison of the VISTA model with different exiting methods with the EyeQ dataset

Method	Accuracy	Precision	Recall	*F*1-score
BRISQUE [22]	0.7692	0.7608	0.7095	0.7112
NBIQA [21]	0.7917	0.7641	0.7509	0.7441
TS-CNN [23]	0.7926	0.7976	0.7446	0.7481
HVS-based [8]	–	0.7404	0.6945	0.6991
MR-CNN [16]	0.8843	0.8697	0.8700	0.8694
DenseNet121-MCF [5]	–	0.8645	0.8497	0.8551
DenseNet121-MCF [5]	0.8722	0.8563	0.8482	0.8506
DenseNet121-RGB [5]	–	0.8194	0.8114	0.815
DenseNet121-RGB [5]	0.8568	0.8481	0.8239	0.8315
ResNet-18-RGB [5]	–	0.804	0.816	0.808
ResNet-18-HSVB [5]	–	0.801	0.816	0.808
ResNet-50-RGBB [5]	–	0.812	0.807	0.810
Resenet-50-HSVB [5]	–	0.770	0.777	0.773
Single-branch SalStructIQA [14]	0.8847	0.8715	0.8645	0.8662
Dual-branch SalStructIQA [14]	0.8897	0.8748	0.8721	0.8723
CNN-RGB [19]	–	0.860	0.862	0.860
CNN combined [19]	–	0.878	0.880	0.878
FGR-Net [15]	0.8947	0.8800	0.8765	0.8782
VISTA	0.9066	0.8843	0.8905	0.8868

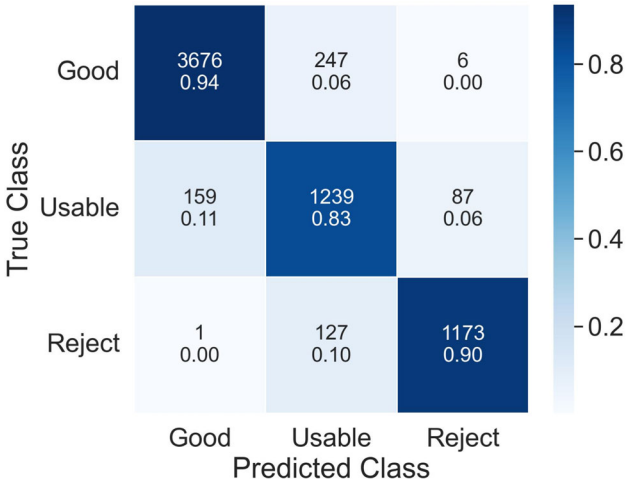


Fig. 7 Confusion matrix for the testing set of the EyeQ dataset with the VISTA model

handcrafted features for quality assessment. In contrast, VISTA is a straightforward model designed for high-resolution images. Our model utilizes only the encoder network and the classifier during testing without extracting prior information from the input fundus images. Notably, VISTA outperforms our previous model, FGR-Net [15], which shares a similar structure but employs low-resolution images of 512×512. This suggests that preserving the original image's size improves the classification rate. In summary, the results underscore the significant improvement achieved by VISTA across all evaluated measures.

To check the scalability and upgradability of the VISTA model on the EyeQ dataset with three classes (*Good*, *Usable*, and *Rejected*), we also computed the confusion matrix with loss functions (MSE) and overall classification accuracy in the test set. Figure 7 shows TPs and TNs of VISTA with a test set of 13, 430 images with loss functions. The model classifies the fundus images into three classes with few mispredictions. For instance, the model with MSE and the first class *Good* classified only six *Rejected* images as *Good* images and 247 *Usable* images as *Good*. This result is intuitive since both *Usable* and *Good* have similar characteristics. Among the different reconstruction losses, MSE yields the highest TP and TN on the test set of the EyeQ dataset.

6.2 Post hoc interpretable results for the VISTA model

The results are illustrated in Figs. 8 and 9, directly stem from the application of attribution methods: Gradient, Grad-CAM, Occlusion, and RISE.

6.2.1 Result analysis

In general terms, the post hoc interpretability algorithms gave similar results regarding the pixels of the image that the model considers for the classification process of a given class.

Specifically focusing on the *Good* class, the results highlighted in Fig. 8 show that some results focus purely on the optic disc and veins. In contrast, others have a preference only for the optic disc (about Grad-CAM and RISE). Also, with RISE, it is clear that, on average, the focus is on the area of the optic disc. In contrast, the Occlusion method yielded more variable and sparse importance maps, possibly due to its reliance on single-patch occlusion. Grad-CAM exhibited a more pronounced impact on the optic disc and veins, with the influence of the crop-and-split operation at the beginning of the pipeline more evident for class *Rejected*, as shown in Fig. 9. Particularly for this class, there is still a notable similarity between the results obtained by RISE and Grad-CAM, but the difference with the Occlusion results has increased significantly. Due to the model's crop-and-split preprocessing, which involves dividing the image into four patches at the beginning of the pipeline, we observed a significant impact compared to the previous choice of Grad-CAM as the state-of-the-art method [15]. This shift in results led us to adopt RISE as a more suitable solution for the current configuration. RISE's ability to generate smooth masks covering multiple patches allowed for a more coherent representation of values of positive importance. These findings underscore the nuanced differences in interpretability outputs, emphasizing the importance of selecting an appropriate method tailored to the characteristics of the model and dataset.

Finally, applying clustering methods to RISE outputs from 150 samples of each class, a notable observation is the pronounced focus on the optic disc region, as Fig. 10 suggests. This concentration reaffirms the clinical significance of the optic disc in fundus image analysis, emphasizing its pivotal role in influencing model predictions.

It is crucial to emphasize the substantial difference between utilizing larger image sizes in training in our enhanced model design, unlike our previous approach [15]. As a result, we observed that it improved both local and global feature extraction facilitated by this updated architecture.

6.2.2 Clustering and global insights

In tandem with individual interpretability methods, we explore techniques that combine local insights to derive a holistic understanding of the fundus images. Specifically,

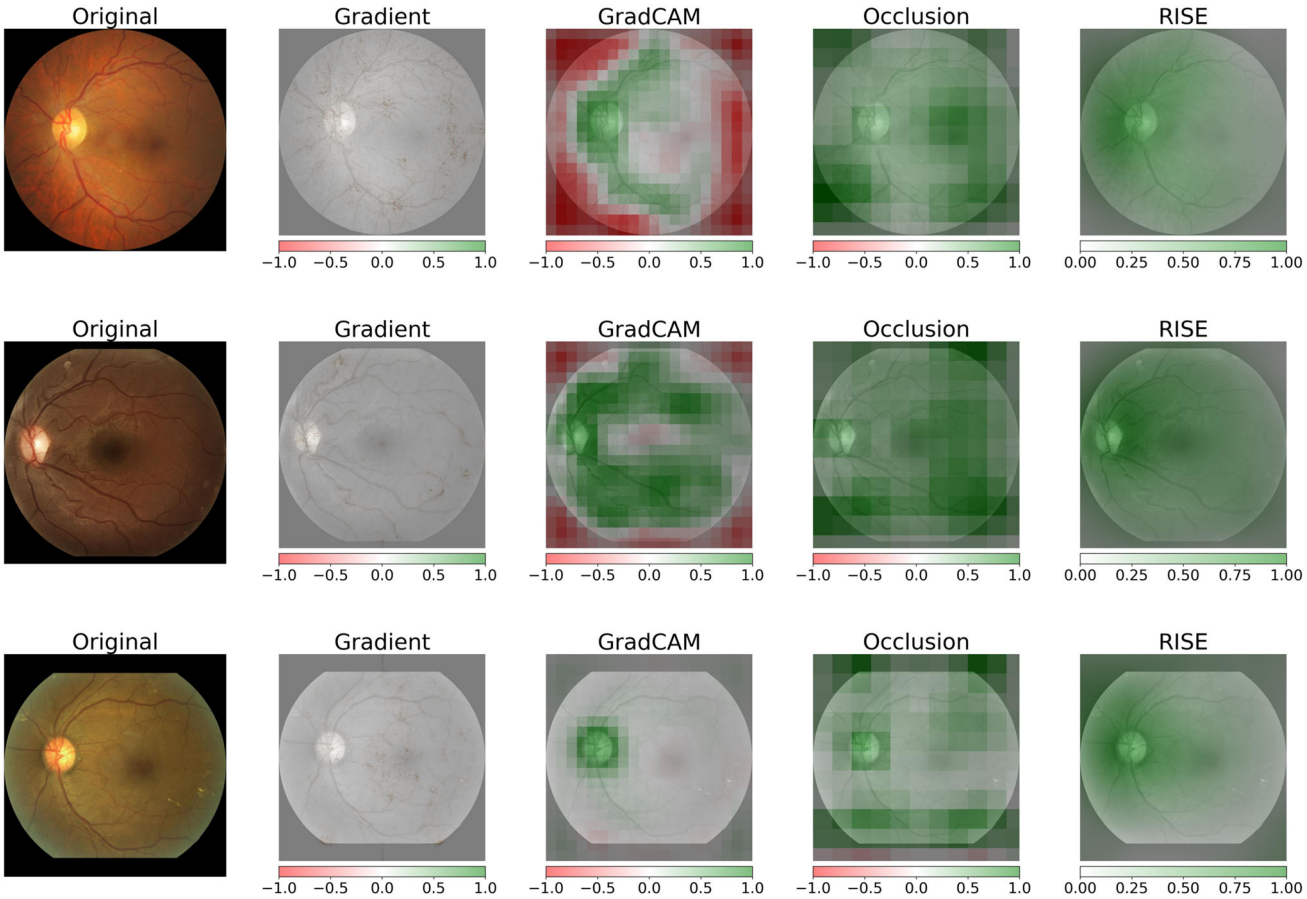


Fig. 8 Comparison of Gradient, Grad-CAM, Occlusion and RISE techniques applied to our model for three class samples *Good*. The gradient method only focuses on low-level features mostly found in blood vessels. Grad-CAM focuses on the optical disc and main vessels but also includes or excludes other areas arbitrarily. Occlusion includes the disc but also various random areas. Finally, RISE improves on all previous methods by focusing on the optic disc and main vessels

we employ the k-means clustering method. This approach captures patterns and groups salient regions, providing a global perspective on the fundus images.

Moreover, the integration of clustering on RISE results yields distinctive patterns, shedding light on localized regions that collectively contribute to the model's decision-making. This comprehensive approach, merging local insights through RISE with global understanding through k-means clustering, enriches our comprehension of the intricate relationships between diverse regions within the fundus images.

6.2.3 Computational time evaluation for real-time feedback

Visualizations highlighting the features on which the model focuses play a crucial role in enabling medical practitioners and technicians to validate the quality of fundus image acquisition. Particularly in the context of mobile fundus photography, real-time feedback on image quality facilitates the capture of high-quality images using cost-effective devices. Considering that the decoder part is not utilized for inference, we assessed the performance of each visualization method on the proposed model. We conducted the tests on GPU NVIDIA Pascal X Titan with 12 GB of RAM-running on an Intel i7 CPU with 32 GB of RAM. The mean computation time over 50 runs was measured, each consisting of a batch with a single sample to simulate real-time conditions. Table 10 outlines the computation time for various methods.[2] The image prediction latency can be used as a baseline, where for the *Good* class, the computation time is 0.18 s (standard deviation: 0.16), and for the *Rejected* class, it is 0.14 s (standard deviation: 0.01).

While RISE outperforms other methods in terms of results because its unique ability to generate

[2] It's essential to highlight that these timings were obtained using a stock PyTorch implementation, with half-precision and no optimizations, including Captum's stock implementation of all interpretability measures, except for RISE, that our own customized implementation was used, which also fulfil the design considerations from Captum [34]; thus, performance enhancement is feasible.

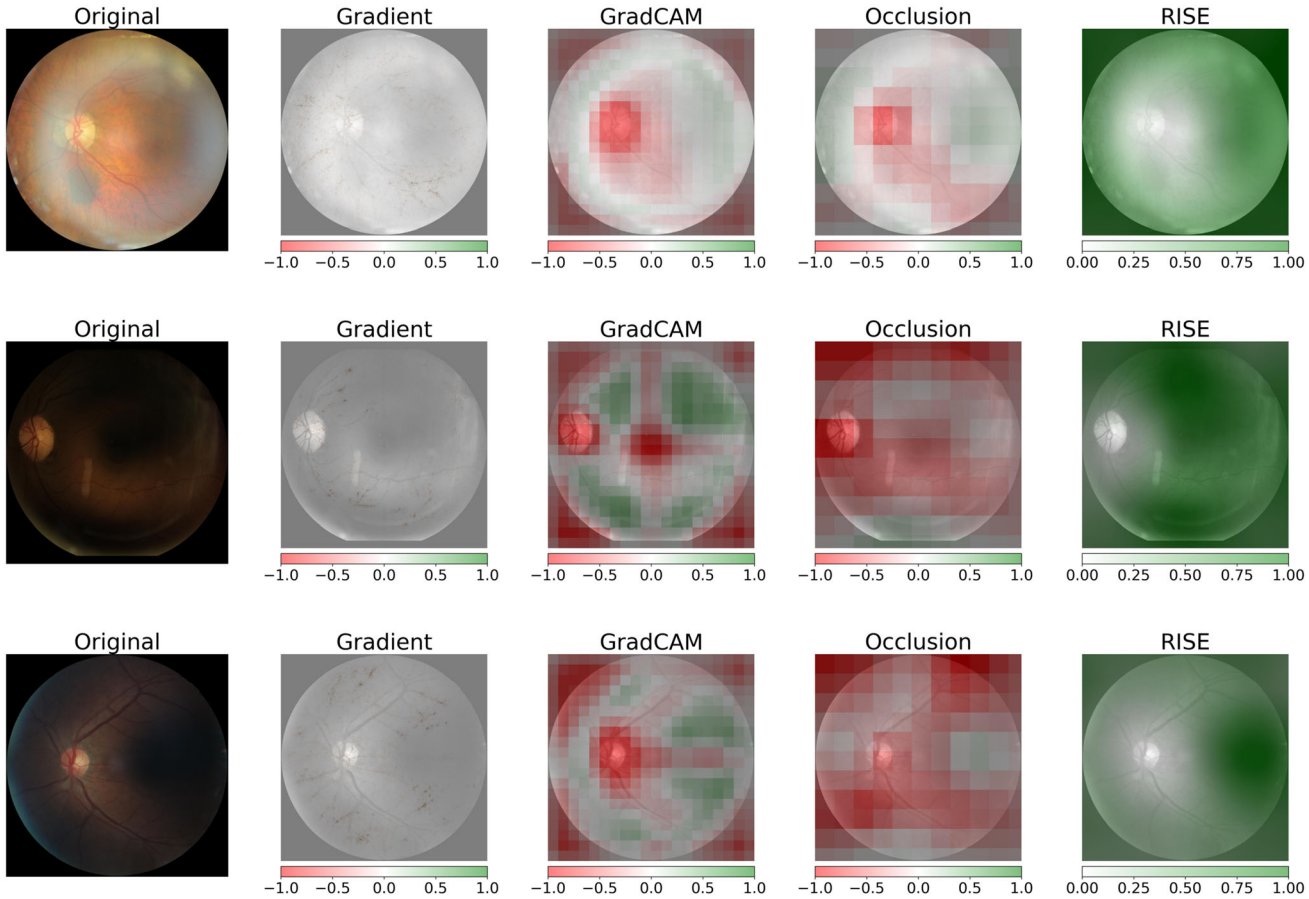


Fig. 9 Comparison of Gradient, Grad-CAM, Occlusion and RISE techniques applied to our model for three class samples *Rejected*. As with class *Good*, RISE provides better results, indicating the crescent-shaped artefact area in the first row and the over-dark areas in the other two rows

comprehensive saliency maps underscores its value in unravelling the intricacies of fundus image quality assessment, it's essential to acknowledge the significant time investment linked to its execution in comparison to other interpretability techniques is crucial. The computational complexity of RISE is prohibitively high (roughly 100 or more times slower than Grad-CAM) because it involves a Monte Carlo integral approximation.

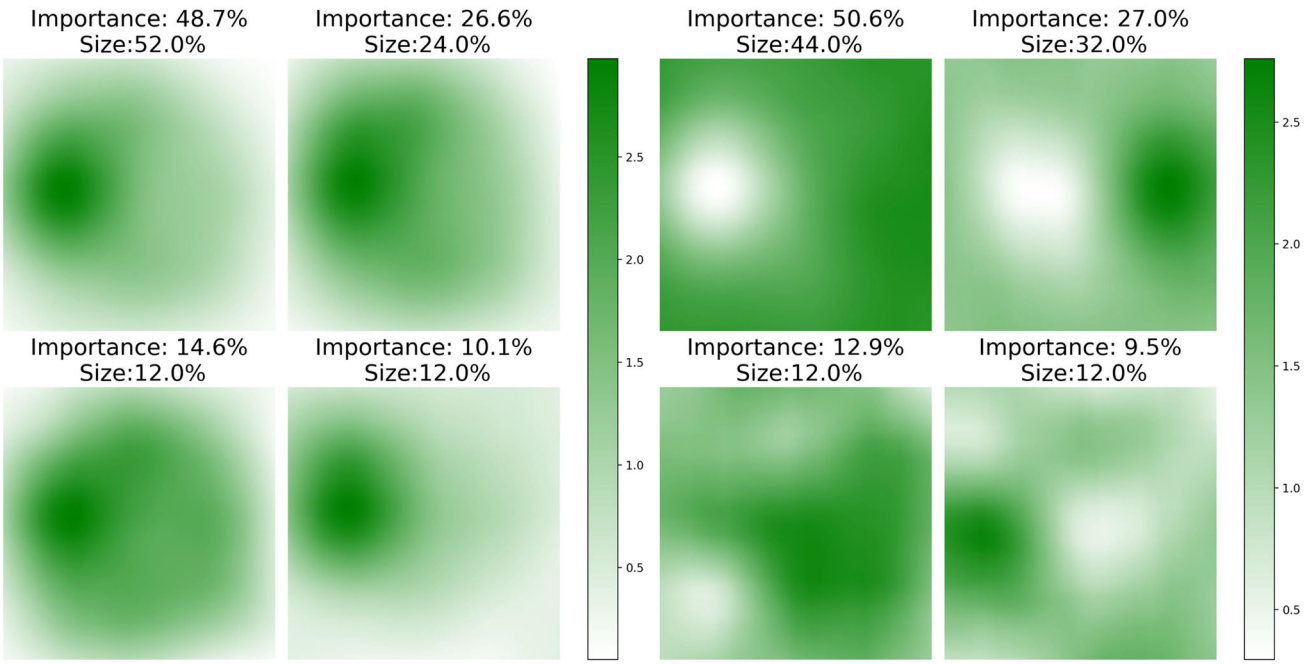
6.3 Intrinsic interpretable results for the VISTA model

The results are shown in Fig. 11, directly arising from incorporating CW into the VISTA architecture for intrinsic interpretation.

6.3.1 Results analysis

In the context of CW, the visualization of activation scores during testing plays a crucial role in understanding the model's decision-making process. Figure 11 shows CW's activation scores at the output layer. Scores are normalized to 0–1 and displayed alongside each image with a corresponding bar, providing insights into concept activation. For the "good" concept, there is minimal confusion with the other two concepts, indicating a clear model response. However, as we shift our focus to the *illumination-bright* and *illumination-dark* concepts, the level of confusion increases. Nevertheless, the model still accurately recognizes these concepts.

The strategic placement of the CW layer within the model architecture is paramount for capturing high-level feature representations; placement in previous layers did not yield a reasonable decomposition. Also, the incorporation of CW's optimization processes maintained a



(a) Cluster for Class **Good**. (b) Cluster for Class **Rejected**.

Fig. 10 Clustering analysis using k-means on RISE outputs, revealing uncovering unique arrangements and underscoring the relevance of the optic disc in enhancing the interpretability of fundus image quality assessment

Table 10 Average time (with standard deviation) in seconds for each interpretation method in different classes. The coefficient of variation measures relative variability relative to the mean

Class	Gradient	Grad-CAM	Occlusion	RISE
Good	0.21 (0.01)	0.25 (0.01)	16.99 (1.04)	414.59 (15.52)
Rejected	0.20 (0.01)	0.24 (0.01)	15.93 (1.27)	392.29 (20.31)

comparable training duration and accuracy with respect to the baseline model, highlighting the efficiency of the implemented methods without imposing a substantial time penalty on the training process. It's well-established that the addition or replacement of CW layers in a model does not significantly impact its classification accuracy [35].

The experiments unveil the intricate interplay between interpretability methods, training strategies, and dataset curation, emphasizing the meticulous considerations essential for robust model development in fundus image quality assessment.

7 Conclusion and future work

This work proposed a deep learning model, VISTA, combining a model for extracting global and local features and autoencoder and multi-layer classifier networks for predicting the gradability of retinal fundus images. The autoencoder consists of two networks: encoder and decoder. The autoencoder network is used to reconstruct the input fundus image. Our model also includes a multi-layer classifier fed by features extracted from the encoder network to rank the gradability of the fundus image as *Good* or *Usable* and *Rejected*. VISTA's learning approach combines the CE loss function based on supervised learning and self-supervised learning by comparing the reconstructed image to the target image (i.e. the input image). The VISTA model based on the VGG-16 backbone as the base of the encoder network and using the MSE as a reconstruction loss function achieved an overall accuracy of 0.91, precision of 0.88, recall of 0.89, and *F*1-score of 0.89. Our model outperformed the state-of-the-art retinal gradability assessment in the three class (*Good*, *Usable*, and *Rejected*) tasks. The VISTA model can correctly identify the visual features of eye image gradability for a more precise grading system. In essence, the integration of

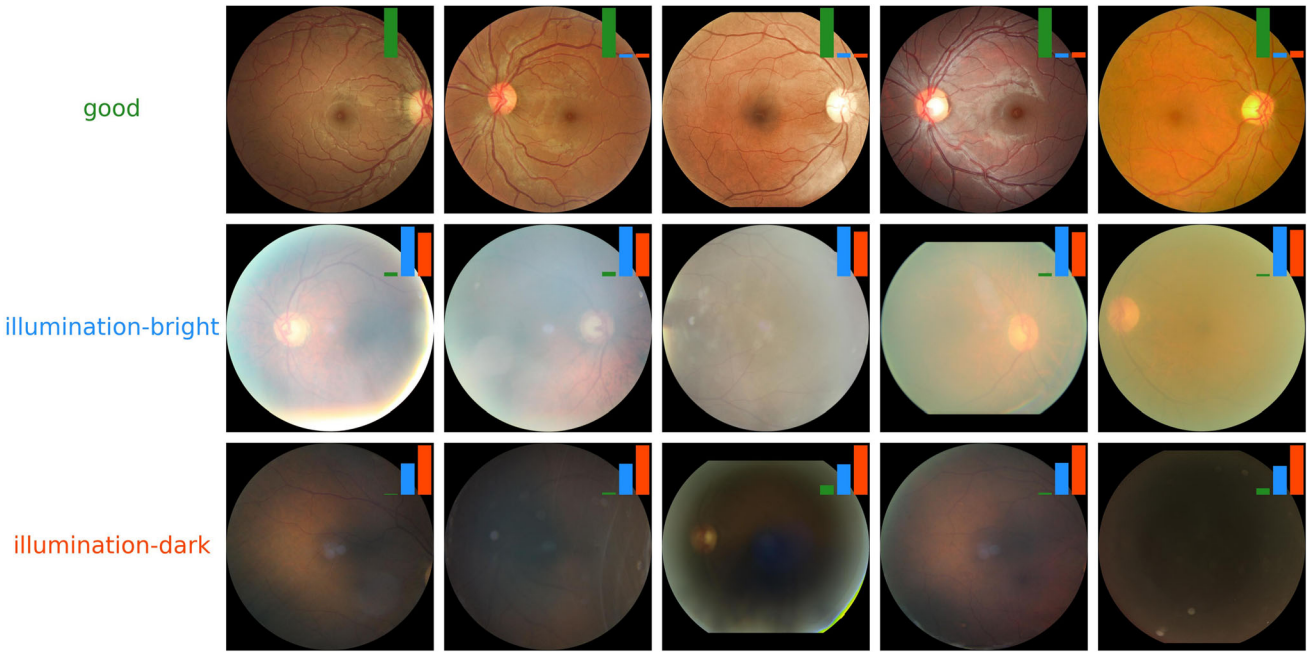


Fig. 11 Top 5 Influential Fundus Images for Annotated Concepts in Concept Whitening. Bar plots alongside each image represent concept activation, with each colour indicating the intensity of the corresponding concept for the particular fundus image

interpretability methods serves as a means to provide ophthalmologists with interpretable visual feedback, elucidating how our model evaluates the quality of fundus images. We will also consider further incorporating interpretability metrics directly into the optimization process to ensure a more holistic approach to model development and refinement. This emphasis on interpretability is pivotal, particularly in medical image analysis, where transparency and trust are primary. Additionally, despite its time-intensive nature, we are proactively engaged in ongoing optimization efforts to enhance the efficiency of RISE. Our goal is to streamline RISE for real-time applications further.

Acknowledgements This work was supported by the research project RetinaReadRisk, funded by EIT Health and Horizon Europe under grant agreement 220718. The authors would like to express their gratitude to the "Rovira i Virgili" for their support.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The dataset used in this study is publicly available and can be accessed at https://github.com/hzfu/EyeQ (Ref. [5]).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Not applicable.

References

1. Khalid S, Abdulwahab S, Rashwan HA, Abdel-Nasser M, Sharaf N, Puig D (2022) Robust yet simple deep learning-based ensemble approach for assessing diabetic retinopathy in fundus images. In: 2022 5th international conference on multimedia, signal processing and communication technologies (IMPACT). IEEE, pp 1–5
2. Jelinek H, Cree MJ (2009) Automated image detection of retinal pathology. CRC Press, Boca Raton
3. Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF (2006) Automated assessment of diabetic retinal image quality based on clarity and field definition. Investig Ophthalmol Vis Sci 47(3):1120–1125
4. MacGillivray TJ, Cameron JR, Zhang Q, El-Medany A, Mulholland C, Sheng Z, Dhillon B, Doubal FN, Foster PJ, Trucco E et al (2015) Suitability of UK biobank retinal images for automatic analysis of morphometric properties of the vasculature. PLoS ONE 10(5):0127914
5. Fu H, Wang B, Shen J, Cui S, Xu Y, Liu J, Shao L (2019) Evaluation of retinal image quality assessment networks in

different color-spaces. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 48–56
6. Lee SC, Wang Y (1999) Automatic retinal image quality assessment and enhancement. In: Medical imaging 1999: image processing, vol 3661. International Society for Optics and Photonics, pp 1581–1590
7. Dias JMP, Oliveira CM, Silva Cruz LA (2014) Retinal image quality assessment using generic image quality indicators. Inf Fusion 19:73–90
8. Wang S, Jin K, Lu H, Cheng C, Ye J, Qian D (2015) Human visual system-based fundus image quality assessment of portable fundus camera photographs. IEEE Trans Med Imaging 35(4):1046–1055
9. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
10. Shen Y, Sheng B, Fang R, Li H, Dai L, Stolte S, Qin J, Jia W, Shen D (2020) Domain-invariant interpretable fundus image quality assessment. Med Image Anal 61:101654
11. Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, Gerven M, Lier R (2018) Explainable and interpretable models in computer vision and machine learning. Springer, Berlin
12. Xu Z, Zou B, Liu Q (2022) A dark and bright channel prior guided deep network for retinal image quality assessment. Biocybern Biomed Eng 42(3):772–783
13. Jiang H, Yang K, Gao M, Zhang D, Ma H, Qian W (2019) An interpretable ensemble deep learning model for diabetic retinopathy disease classification. In: 2019 41st annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp 2045–2048
14. Xu Z, Zou B, Liu Q (2023) A deep retinal image quality assessment network with salient structure priors. Multimed Tools Appl 82:34005–34028
15. Khalid S, Rashwan HA, Abdulwahab S, Abdel-Nasser M, Quiroga FM, Puig D (2024) FGR-Net: interpretable fundus image gradeability classification based on deep reconstruction learning. Expert Syst Appl 238:121644
16. Raj A, Shah NA, Tiwari AK, Martini MG (2020) Multivariate regression-based convolutional neural network model for fundus image quality assessment. IEEE Access 8:57810–57821
17. Li Q, Wei H, Hua D, Wang J, Yang J (2024) Stabilization of semi-Markovian jumping uncertain complex-valued networks with time-varying delay: a sliding-mode control approach. Neural Process Lett 56(2):1–22
18. Li Q, Liang J, Gong W, Wang K, Wang J (2024) Nonfragile state estimation for semi-Markovian switching CVNS with general uncertain transition rates: An event-triggered scheme. Math Comput Simul 218:204–222
19. Muddamsetty SM, Moeslund TB (2021) Multi-level quality assessment of retinal fundus images using deep convolution neural networks. In: 16th international joint conference on computer vision, imaging and computer graphics theory and application. SCITEPRESS Digital Library, pp 661–668
20. Li S, Wang M, Hou C (2019) No-reference stereoscopic image quality assessment based on shuffle-convolutional neural network. In: 2019 IEEE visual communications and image processing (VCIP). IEEE, pp 1–4
21. Ou F-Z, Wang Y-G, Zhu G (2019) A novel blind image quality assessment method based on refined natural scene statistics. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 1004–1008
22. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
23. Yan Q, Gong D, Zhang Y (2018) Two-stream convolutional networks for blind image quality assessment. IEEE Trans Image Process 28(5):2200–2211
24. Pérez AD, Perdomo O, González FA (2020) A lightweight deep learning model for mobile eye fundus image quality assessment. In: 15th international symposium on medical information processing and analysis, vol 11330. SPIE, pp 151–158
25. Zhou X, Wu Y, Xia Y (2020) Retinal image quality assessment via specific structures segmentation. In: Ophthalmic Medical Image Analysis: 7th international workshop, OMIA 2020, held in conjunction with MICCAI 2020, Lima, Peru, 8 Oct 2020, Proceedings 7. Springer, pp 53–61
26. Liu Y-P, Lv Y, Li Z, Li J, Liu Y, Chen P, Liang R (2021) Blood vessel and background separation for retinal image quality assessment. IET Image Proc 15(11):2559–2571
27. Chen Z, Huang L (2022) Deep convolutional neural network for image quality assessment and diabetic retinopathy grading. MICCAI challenge on mitosis domain generalization. Springer, Cham, pp 31–37
28. Zago GT, Andreão RV, Dorizzi B, Salles EOT (2018) Retinal image quality assessment using deep learning. Comput Biol Med 103:64–70
29. Zhang F, Xu X, Xiao Z, Wu J, Geng L, Wang W, Liu Y (2020) Automated quality classification of colour fundus images based on a modified residual dense block network. Signal Image Video Process 14:215–223
30. Hou J, Lin W, Zhao B (2020) Content-dependency reduction with multi-task learning in blind stitched panoramic image quality assessment. In: 2020 IEEE international conference on image processing (ICIP). IEEE, pp 3463–3467
31. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
32. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, 6–12 Sept 2014, Proceedings, Part I 13. Springer, pp 818–833
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
34. Stanchi O, Ronchetti F, Quiroga F (2023) The implementation of the rise algorithm for the captum framework. In: Conference on cloud computing, big data & emerging topics. Springer, pp 91–104
35. Chen Z, Bei Y, Rudin C (2020) Concept whitening for interpretable image recognition. Nat Mach Intell 2(12):772–782
36. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2:: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
37. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
38. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
39. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In:

Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500
40. Dai Z, Liu H, Le QV, Tan M (2021) CoAtNet: marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst 34:3965–3977
41. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 31
42. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556