Searching Discriminative Regions for Convolutional Neural Networks in Fundus Image Classification With Genetic Algorithms

Yibiao Rong, Tian Lin, Haoyu Chen, Zhun Fan, *Senior Member, IEEE*, and Xinjian Chen, *Senior Member, IEEE*

***Abstract*— Deep convolutional neural networks (CNNs) have been widely used for fundus image classification and have achieved very impressive performance. However, the explainability of CNNs is poor because of their black-box nature, which limits their application in clinical practice. In this paper, we propose a novel method to search for discriminative regions to increase the confidence of CNNs in the classification of features in specific category, thereby helping users understand which regions in an image are important for a CNN to make a particular prediction. In the proposed method, a set of superpixels is selected in an evolutionary process, such that discriminative regions can be found automatically. Many experiments are conducted to verify the effectiveness of the proposed method. The average drop and average increase obtained with the proposed method are 0 and 77.8%, respectively, in fundus image classification, indicating that the proposed method is very effective in identifying discriminative regions. Additionally, several interesting findings are reported: 1) Some superpixels, which contain the evidence used by humans to make a certain decision in practice, can be identified as discriminative regions via the proposed method; 2) The superpixels identified as discriminative regions are distributed in different locations in an image rather than focusing on regions with a specific instance; and 3) The number of discriminative superpixels obtained via the proposed method is relatively small. In other words, a CNN model can employ a small portion of the pixels in an image to increase the confidence for a specific category.**

Received 25 September 2023; revised 21 April 2024, 8 July 2024, 7 August 2024, and 17 September 2024; accepted 6 October 2024. Date of publication 16 October 2024; date of current version 22 October 2024. This work was supported in part by Guangdong Natural Science Foundation under Grant 2022A1515011396, in part by the National Key Research and Development Program of China under Grant 2021ZD0111502, and in part by the Science Research Startup Foundation of Shantou University under Grant NTF20021. The associate editor coordinating the review of this article and approving it for publication was Dr. Laura Boucheron. *(Corresponding authors: Yibiao Rong; Zhun Fan; Xinjian Chen.)*

Yibiao Rong is with the College of Engineering, Shantou University, Shantou 515063, China, and also with the Artificial Intelligence and Modern Ultrasonic Engineering Technology Research Center of Guangdong Province, Shantou 515063, China (e-mail: ybrong@stu.edu.cn).

Tian Lin and Haoyu Chen are with the Joint Shantou International Eye Center, Shantou University, Shantou 515063, China, and The Chinese University of Hong Kong, Shantou 515041, China (e-mail: drtianlin@163.com; drchenhaoyu@gmail.com).

Zhun Fan was with the College of Engineering, Shantou University, Shantou 515063, China. He is now with Shenzhen Institute for Advanced Study, University of Electronic Science and Technology, Shenzhen 518000, China (e-mail: zfan@stu.edu.cn).

Xinjian Chen is with the School of Electrical and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: xjchen@suda.edu.cn).

Digital Object Identifier 10.1109/TIP.2024.3477932

***Index Terms*— Discriminative regions, genetic algorithms, convolutional neural networks, fundus image classification.**

I. Introduction

FUNDUS images are effective tools for observing the progression of different eye diseases, such as diabetic retinopathy (DR) and glaucoma. The interesting clinical features of an eye, such as the optic disk and vessels, can also be clearly observed in a fundus image [1]. The images in Fig. 1 are examples of different fundus images, where Fig. 1(a) is a fundus image from a normal subject, Fig. 1(b) is a fundus image from a subject with DR, and Fig. 1(c) is a fundus image from a subject with glaucoma. The fundus images from subjects with different diseases exhibit different visual features. For example, there are lesions, with hard exudates and soft exudates, in fundus images from subjects with DR. In fundus images of subjects with glaucoma, the ratio of the size of the optic disk to the size of the optic cup is generally abnormal [2]. These visual features are evidence used by ophthalmologists to make a diagnosis in clinical practice.

As the number of patients with eye diseases increases, there is a desire to develop a computer-aided diagnosis system that can automatically and accurately identify diseases on the basis of fundus images [3], [4]. Numerous methods have been proposed for developing computer-aided systems for detecting eye diseases, including the use of classical image processing techniques, e.g., fuzzy methods [5], to detect related eye diseases. With the revival of deep learning [6], especially deep convolutional neural networks (CNNs) [7], many researchers have employed deep learning techniques to develop computer-aided diagnosis systems. For example, Ting et al. [8] developed a deep learning system for recognizing DR and related eye diseases via fundus images. Raghaverndra et al. [1] employed CNNs for the accurate diagnosis of glaucoma on the basis of fundus images. Cen et al. [9] also employed deep neural networks to automatically detect different diseases on the basis of fundus images.

A. Techniques for Explainability

Although the precision achieved by CNNs for fundus image classification is very impressive, explaining why CNNs make a specific prediction given a particular fundus image is difficult

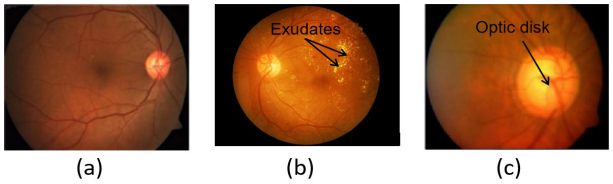


Fig. 1. Examples of different fundus images. (a) Normal. (b) DR. (c) Glaucoma.

because of the black-box nature of CNNs [10]. Notably, explainability for CNNs is very important, especially when CNNs are applied for high-impact and high-risk tasks, such as medical diagnosis [11]. The research area of explainable deep learning (also termed XAI) [12] has drawn widespread attention. Many methods have been proposed to explain why CNNs make a specific prediction, among which saliency map-based methods [13] have been widely studied, with the purpose of generating a saliency map to highlight the regions in an image that are important for a CNN to make a specific prediction [14].

Saliency map-based methods can be divided into class activation map (CAM)-based methods [15] and perturbation-based methods [16]. The CAM was proposed by Zhou et al. [15], who defined a CAM as a discriminative region used in CNNs to identify a specific category of feature and generated by mapping the predicted class score back to the previous convolutional layer. Inspired by the original CAM definition [15], several improved methods based on CAMs have been proposed, such as Grad-CAM [17], Grad-CAM++ [18], Score-CAM [19], Ablation-CAM [20], Eigen-CAM [21], Relevance-CAM [22] and Shap-CAM [23].

Unlike the mechanism for which backpropagation is used to generate discriminative regions in CAM-based methods, in perturbation-based methods, a part of the input for a CNN model is first modified. The change in output is then observed. The degree of the change in output indicates which parts of the input are particularly important for a CNN to make a prediction [16], [24]. Local interpretable model-agnostic explanation (LIME) [25] is a representative perturbation-based method that employs the occlusions of superpixels and linear models to obtain saliency maps. Several improved methods have been proposed on the basis of the framework of LIME. Ribeiro et al. [26] improved LIME by maximizing the coverage region of an explanation. Wang et al. [27] proposed a method termed the multiobjective evolutionary computation-based model-agnostic method (MO-LIME) to overcome the limitations of LIME, such as the expensive sampling process and the prefixed number of interpretable features.

There are also other forms of perturbation-based methods. For example, Petsiuk et al. [28] estimated the importance of pixels by dimming them in random combinations. Fong and Vedaldi [14] formulated the problem of searching for discriminative regions as an information maximization problem and solved the problem by using a local search with a gradient descent method. Dabkowski and Gal [29] proposed a fast saliency detection method and trained a model to predict a saliency map with a single feed-forwards pass. Chang et al. [30] explained image classifiers via counterfactual generation. They sampled plausible image in-fills by conditioning a generative model and then searched the image regions that changed the decision of the classifier the most after in-fill addition. Del Ser et al. [31] presented a framework for generating counterfactual explanations on the basis of the argument that trust can be achieved via counterfactual explanations on the basis of the hypothetical input conditions under which the output changes.

In addition to CAM-based methods and perturbation-based methods, many other methods for assessing CNNs have been proposed. Zeiler and Fergus [32] introduced a visualization technique to provide insight into the functions of intermediate layers and the operation of models. Kuo [33] proposed a mathematical model, termed rectified correlations on a sphere, for use with CNNs. Shang et al. [34] proposed a concatenated rectified linear unit to improve CNNs. Wang et al. [35] proposed a method with interactive visualization to explain the behaviours of CNNs. Xuan et al. [36] proposed a visual system for comparative studies of CNNs.

B. XAI in Fundus Image Classification

The above methods were developed primarily for explaining CNNs in natural image classification. Some existing works have explored whether existing methods, such as CAM-based methods and perturbation-based methods, can provide valid explanations for CNNs in fundus image classification. For example, Cen et al. [9] employed CAM-based methods to generate heatmaps and indicate important regions for CNNs in fundus image classification. Additionally, Deperlioglu et al. [37] employed CAM-based methods to identify important regions for CNNs in glaucoma diagnosis. Chang et al. [38] employed adversarial examples to explain the rationale of the decisions made by a deep learning model in glaucoma classification. Kamal et al. [39] used an adaptive neuro-fuzzy inference system and a pixel density analysis method to provide explanations for models in glaucoma prediction.

Although existing works have made some progress in the area of explainable deep learning in fundus image classification, the issue of understanding the behaviours of CNNs remains largely an open problem. In this paper, we propose a novel method to search for discriminative regions for CNNs in fundus image classification based on genetic algorithms, thereby helping users understand which regions in an image are important for a CNN model to make a specific prediction. The proposed method follows the general steps of perturbation-based methods [25], where discriminative regions are obtained by modifying the inputs of CNNs and observing the changes in the outputs. To accelerate the process of searching for discriminative regions, in the proposed method, a set of superpixels is involved in an evolutionary process, such that a combination of the superpixels (or discriminative regions) that can increase the confidence of CNNs for a specific category can be automatically identified.

Although the proposed method follows the general steps of perturbation-based methods [25], there are fundamental differences in the approach used to search for discriminative regions compared with the existing perturbation-based

TABLE I
THE CHARACTERISTICS OF DIFFERENT METHODS

Methods	Methodology	Intrinsic	Post-hoc	Model-Agnostic	Model-Specific	Global	Local
GradCAM [17]	Utilizing the gradients of the target concepts flowing into the final convolutional layer to produce a coarse localization map highlighting important regions.	×	√	√	×	×	√
GradCAM++ [18]	Using a weighted combination of the positive partial derivatives of the convolutional layer feature maps of the previous layer with respect to a specific class score as weights to generate a visual explanation .	×	√	√	×	×	√
ScoreCAM [19]	Generating the visual explanation via a linear combination of activation maps and score-based weights extracted based on level of confidence.	×	√	√	×	×	√
Rise [28]	Estimating the importance of pixels by dimming them in random combinations.	×	√	√	×	×	√
Mask [14]	Solving an information maximization problem using a local search via gradient descent methods.	×	√	√	×	×	√
MO-LIME [27]	Formulating the problem as a multi-objective optimization problem. Solving the problem with NSGA-II.	×	√	√	×	×	√
Proposed	Formulating the problem as a combinatorial optimization problem. Solving the problem with genetic algorithms.	×	√	√	×	×	√

methods. To clarify the differences among different methods, the characteristics of different methods, including CAM-based methods [17], [18], [19], perturbation-based methods [14], [27], [28], etc., according to different criteria are given, as summarized in Table I, which includes the details for each method in terms of methodology, structure (intrinsic or post hoc), scope (local or global) and dependent criteria (model specific or model agnostic) [40]. With respect to structure, if XAI models can be integrated as a part of a network, they are considered intrinsic. If XAI models are used to explain the networks, they are post hoc. With respect to scope, if XAI models can access model data and provide explanations, they are local. If XAI models can access data and provide feedback regarding network behaviours, they are global. Local models access individual instances, whereas in global models, the network architecture is treated as a black box. With respect to the dependent criteria, some XAI models are designed to work with specific AI systems (i.e., model specific), whereas other models can be generalized across several networks (i.e., model agnostic).

CAM-based methods [17], [18], [19] and perturbation-based methods [14], [27], [28] share similar characteristics in terms of structure, scope, and dependent criteria and are post hoc, local, and model agnostic. The primary differences among different approaches are the methods used for generating discriminative regions. CAM-based methods [17], [18], [19] employ backpropagation to generate visual explanations. Perturbation-based methods explore various ways to modify the inputs for CNNs to obtain discriminative regions. For example, Rise [28] estimated the importance of pixels by dimming them in random combinations. Mask [14] obtained discriminative regions via a local search with gradient descent methods to solve an optimization problem. MO-LIME [27] employed a multiobjective algorithm to search for discriminative regions. In the proposed method, the problem of searching for discriminative regions is formulated as a combinatorial optimization problem and solved via genetic algorithms.

Notably, in both the proposed method and MO-LIME [27], superpixels in evolutionary algorithms are used to search for discriminative regions. However, there are differences between the proposed method and MO-LIME, such as the number of objectives, the ways used to achieve the objectives and the forms of individuals. Specifically, in MO-LIME, there are two objectives: maximizing the probability of CNNs predicting a local explanation as a specific class label and minimizing the number of selected superpixels. In the proposed method, there is only one objective, which is to maximize the increase in the confidence of CNNs for a specific category. Experiments show that although the minimization of the number of selected superpixels is not employed as an objective in the proposed method, the number of discriminative pixels obtained by the proposed method is often relatively small, with the ratio to the total number of pixels in an image generally being less than 0.25.

Since the objectives for the evolutionary algorithms are different, the methods used to achieve the objectives also vary. In [27], nondominated sorting genetic algorithm II (NSGA-II) [41] was employed to evolve local explanations, and a set of nondominated solutions was returned. In the proposed method, genetic algorithms [42], which consist of fitness evaluation, selection, crossover and mutation, are employed to evolve the discriminative regions, and the individual with the best fitness value is returned. In addition, in MO-LIME, each dimension of an individual is a real value within the range of 0 to 1; this approach increases the search space and reduces efficiency. In the proposed method, the individuals are directly involved in the evolutionary process in the form of binary vectors. In addition, the methods used for crossover and mutation for the individuals in the proposed method are different from those in MO-LIME. The average drop and average increase obtained with the proposed method are 2.97% and 84.5%, respectively, which are significantly better than those (16.4% and 35.5%) obtained with MO-LIME. Notably, the differences between the proposed method and MO-LIME could lead to a significant improvement in the search for discriminative regions and increase the confidence of CNNs in the classification of features in a specific category.

Although the proposed method was originally designed to search for discriminative regions in CNN-based fundus image classification, it can also be applied to search for discriminative regions in other CNN-based image classification tasks (see Section III-C). We conducted a series of experiments to verify the effectiveness of the proposed method, and in the process, we obtained some interesting findings, which are summarized as follows:

1) Some superpixels, which contain evidence used by humans to make a certain decision in practice, can be identified as discriminative regions via the proposed method. For

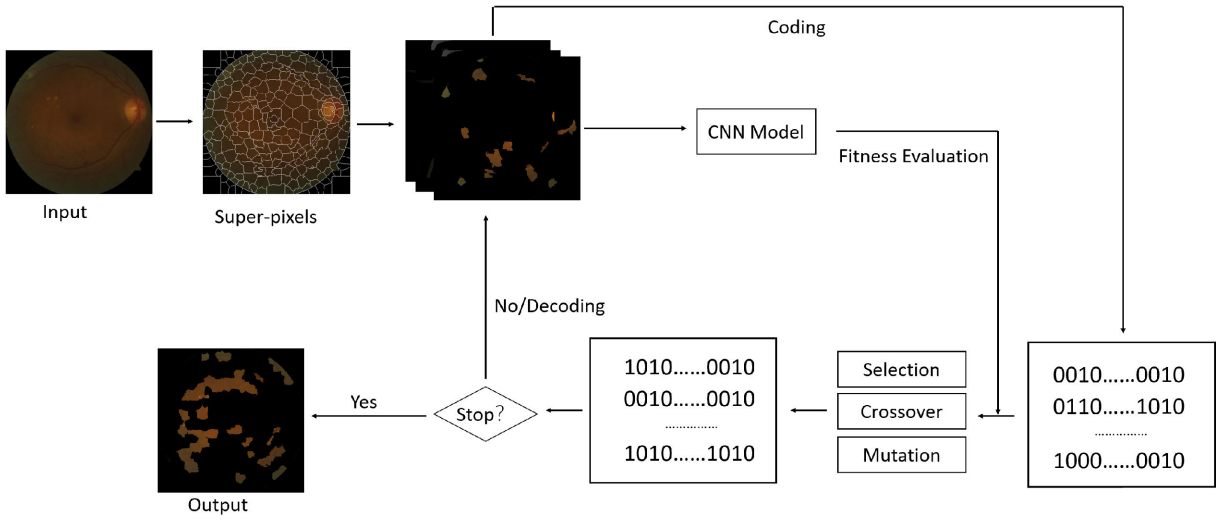


Fig. 2. The flowchart of the proposed method. A set of superpixels is first generated for a given image. A population with a specific number of individuals is then randomly initialized for the superpixels. The fitness value of each individual is computed via the trained CNN model. Evolutionary operations, such as selection, crossover and mutation, are employed to evolve the individuals. Once new individuals are generated, the fitness value of each new individual is computed. The evolutionary process is repeated until the stopping criteria are met.

example, in the case of fundus image classification with DR, the proposed method can search for some superpixels with lesions, which are types of evidence used for the diagnosis of DR in clinical practice, as discriminative regions.

2) The superpixels identified as discriminative regions are distributed in different locations in an image rather than focusing on regions with a specific instance. In other words, even though some superpixels do not contain the evidence used by humans to make a certain decision in practice, they can be identified as discriminative regions with the proposed method.

3) The number of discriminative pixels obtained with the proposed method is relatively small (the ratio to the total number of pixels in an image is less than 0.25), even though the minimization of the number of selected superpixels is not employed as an objective in the proposed method. In other words, a CNN model can employ a small portion of the pixels in an image to increase the confidence for a specific category.

II. Methods

The flowchart of the proposed method is shown in Fig. 2. Given an image x and a trained CNN model $f(\cdot)$, a set of superpixels is first generated for image x. A population with N individuals is then randomly initialized for the superpixels. The fitness value of each individual is computed via the CNN model $f(\cdot)$. T rounds of evolution are performed for the individuals. Each round of evolution involves three evolutionary operations: selection, crossover and mutation. After new individuals are generated, the fitness value of each new individual is computed, and a new round of evolution is performed. Algorithm 1 summarizes the evolutionary process. In Algorithm 1, the subscripts in $Vmask_{t,i}$ indicate that individual $Vmask$ is the i-th individual in the t-th generation. For simplicity, in the following introduction, we omit the subscript t or i.

Algorithm 1 Evolutionary Process

Given: N: Population size. T: The maximum number of generations. p_c: The probability of crossover. p_m: The probability of mutation. S: The number of super-pixels. M: The number of locations for mutation. x: An image. $f(\cdot)$: A trained CNN model.

Initialization: generating a set of randomized individuals $\{Vmask_{0,i}\}_{i=1}^{N}$, and computing their fitness values;

For $t = 1$ to T

Selection: selecting N individuals from $\{Vmask_{t-1,i}\}_{i=1}^{N}$ to generate a new generation $\{Vmask_{t,i}\}_{i=1}^{N}$;

Crossover: for each pair $\{Vmask_{t,2n-1}, Vmask_{t,2n}\}_{n=1}^{\lfloor N/2 \rfloor}$, performing crossover with probability p_c;

Mutation: for each individual after crossover, performing mutation with probability p_m;

Evaluation: computing the fitness value for each new individual.

Endfor

Return: The individual with the best fitness value.

A. Coding & Decoding

Given an image x, the first step of the proposed method is to generate superpixels for image x. Genetic algorithms are then employed to identify a combination of superpixels that can increase the confidence of a CNN for a specific category. In this work, the linear iterative clustering algorithm, also known as the SLIC algorithm [43], is utilized to generate superpixels. In the evolutionary process, the superpixels are encoded into a set of binary vectors with a length of S, where S is the number of superpixels in image x. Each binary vector is termed an individual and is used to determine which superpixels are preserved and which are removed. Each dimension in a binary vector corresponds to a superpixel. If the value in the corresponding dimension for a superpixel is 1, then this superpixel is preserved; otherwise, it is removed.

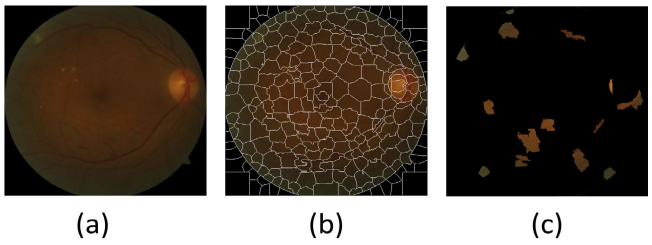


Fig. 3. An example demonstrating the process of coding and decoding. (a) An original image. (b) The image with superpixels. (c) The decoded result.

Fig. 3 is an example used to demonstrate the process of coding and decoding, where Fig. 3(a) is a fundus image, Fig. 3(b) is an image with superpixels, and Fig. 3(c) is a decoded result based on a binary vector. In the following introduction, we employ the symbols $Vmask$ to represent a binary vector for the superpixels and $mask$ to represent the corresponding decoded result.

B. Initialization

Once the coding and decoding strategies are determined, the individuals can be initialized. In this paper, an individual is a binary vector with a length of S used for determining which superpixels are preserved for CNNs to make a prediction. We represent an individual as $Vmask_i$, $i = 1, 2, 3, \ldots, N$, where N is the number of individuals in a generation. To initialize an individual $Vmask_i$, the values in $Vmask_i$ are randomly drawn from a standard uniform distribution in the open interval $(0, 1)$. Thresholding is then performed on $Vmask_i$, where the values greater than a threshold are set to 1 and the others are set to 0. During the thresholding process, the threshold value influences the number of superpixels preserved in the initialization step. Increasing the threshold value results in fewer preserved superpixels. However, it has minimal impact on the evolutionary algorithm when searching for individuals with the best fitness values (see Section III-A). In this paper, we set the threshold value to 0.9.

C. Fitness Evaluation

In the fitness evaluation step, an individual is decoded for a corresponding image via the strategy introduced in the coding and decoding section. The decoded image is then input into a CNN to make a prediction. It is assumed that $Y^c = f(x)$, where Y^c represents the predicted score for category c when the original image x is used as an input for a CNN. Additionally, $O^c = f(mask)$, where O^c represents the predicted score for category c when the decoded image is used as an input. The fitness value for an individual is then defined as $fit = O^c - Y^c$. Since the purpose of this work is to search for discriminative regions that can increase the confidence of a CNN for a specific category of features in classification, a large fitness value indicates a high-quality individual.

D. Selection

During selection, two individuals, $Vmask_i$ and $Vmask_j$, are randomly selected from the current generation. If the fitness value of $Vmask_i$ is greater than or equal to the fitness value of $Vmask_j$, then individual $Vmask_i$ will be selected for the next generation. Otherwise, $Vmask_j$ will be selected. This process is repeated N times to ensure that the number of individuals in the next generation is equal to N.

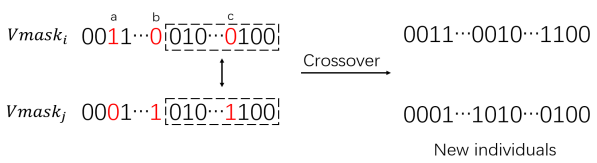


Fig. 4. An example of the crossover process. The first step in crossover is to identify the locations where the values of $Vmask_i$ and $Vmask_j$ are different. In the given example, the values of $Vmask_i$ and $Vmask_j$ at locations a, b, and c are different and are labelled in red. The second step is to randomly select a location from the locations where the values of $Vmask_i$ and $Vmask_j$ are different for crossover. In the given example, location b is selected; thus, the bits associated with location b in $Vmask_i$ and $Vmask_j$ are interchanged to generate new individuals.

TABLE II
IMPLEMENTATION DETAILS

Step	Implementation Details
Initialization	The value for each dimension of an individual is randomly sampled from (0, 1) and converted to 0 or 1 by thresholding.
Evaluation	Decoding an individual into an image. The increase in confidence $O^c - Y^c$ is used as the fitness value (fit) to evaluate each individual.
Selection	Selecting two individuals randomly, if $fit_i \geq fit_j$, $Vmask_i$ is selected; otherwise, $Vmask_j$ is selected. This process is repeated N times.
Crossover	For each pair of individuals $\{Vmask_{2n-1}, Vmask_{2n}\}_{n=1}^{\lfloor N/2 \rfloor}$, crossover is performed with probability p_c.
Mutation	M locations are randomly selected for an individual. Mutation is performed at each selected location with probability p_m.

E. Crossover

Given two individuals $Vmask_i$ and $Vmask_j$, to ensure the crossover between $Vmask_i$ and $Vmask_j$ to generate new individuals, in the proposed method, the location for crossover is randomly selected from the locations where the values of $Vmask_i$ and $Vmask_j$ are different. Fig. 4 shows an example of the crossover process, where the values at locations a, b, and c in $Vmask_i$ and $Vmask_j$ are different. The location for crossover is randomly selected from a, b, or c. In Fig. 4, location b is selected for crossover; thus, the bits associated with location b in $Vmask_i$ and $Vmask_j$ are interchanged with a probability of p_c.

F. Mutation

The purpose of mutation is also to generate new individuals. In this work, $M (M \leq S)$ locations for a given individual are randomly selected for mutation. The probability of mutation at each selected location is p_m. Since the value of an individual $Vmask_i$ is either zero or one, the mutation process at a location involves changing the value to either zero or one depending on the original value at the location. If the original value at a location is zero, it could be changed to one after mutation. If the original value is one, it could be changed to zero after mutation. In summary, Table II provides the implementation details for each step.

III. EXPERIMENTS

A total of 6132 fundus images were collected from Joint Shantou International Eye Center, Shantou University and the

TABLE III
THE STRUCTURE OF THE CNN APPLIED IN THIS WORK

Layer	Input	Insize	K	S	Outsize
conv1	image	$512 \times 512 \times 3$	3×3	2	$255 \times 255 \times 32$
conv2	conv1	$255 \times 255 \times 32$	3×3	1	$253 \times 253 \times 32$
pool1	conv2	$253 \times 253 \times 32$	3×3	2	$126 \times 126 \times 32$
conv3	pool1	$126 \times 126 \times 32$	3×3	2	$62 \times 62 \times 64$
conv4	conv3	$62 \times 62 \times 64$	3×3	1	$60 \times 60 \times 64$
pool2	conv4	$60 \times 60 \times 64$	3×3	2	$29 \times 29 \times 64$
conv5	pool2	$29 \times 29 \times 64$	3×3	2	$14 \times 14 \times 128$
conv6	conv5	$14 \times 14 \times 128$	3×3	1	$12 \times 12 \times 128$
pool3	conv6	$12 \times 12 \times 128$	3×3	2	$5 \times 5 \times 128$
conv7	pool3	$5 \times 5 \times 128$	3×3	1	$3 \times 3 \times 256$
conv8	conv7	$3 \times 3 \times 256$	3×3	1	$1 \times 1 \times 256$
fc1	conv8	$1 \times 1 \times 256$	1×1	1	$1 \times 1 \times 512$
fc2	fc1	$1 \times 1 \times 512$	1×1	1	$1 \times 1 \times 512$
fc3	fc2	$1 \times 1 \times 512$	1×1	1	$1 \times 1 \times 3$

K–kernel size, S–stride, conv–convolutional layer, pool– pooling layer, fc–fully connected layer.

Chinese University of Hong Kong for the experiments. The images were classified into three categories: DR images, glaucoma images and normal images. There were 2152 DR images, 1303 glaucoma images, and 2677 normal images selected. These images were transformed to a size of 512×512 and divided into a training set, a validation set and a test set at a ratio of $8 : 1 : 1$.

Table III summarizes the network structure applied in this work. Note that each layer in the network is followed by a ReLU layer, which is not shown in Table III. The experiments presented in this paper were conducted via MATLAB. The toolbox MatConvNet [44] was employed to train the CNNs. The loss function for the CNNs was Softmax loss. CNN training was performed over 41 epochs, with learning rates of 0.002 (1 to 11 epochs), 0.0002 (12 to 27 epochs) and 0.00002 (28 to 41 epochs). Once training was completed, the accuracy obtained with the trained CNN for the test set was 0.89.

The parameters of genetic algorithms include the size of the population N, the probability of crossover p_c, the probability of mutation p_m, the number of locations for mutation M, the number of superpixels S and the maximum number of iterations T, among others. These parameters affect the performance of the proposed method. For example, if the number of superpixels S is too large (e.g., equal to the number of pixels in an image), the complexity of the discriminative regions will be very high. If the number of superpixels S is too small (e.g., equal to 1), the combinations of superpixels will be limited, and the genetic algorithms may be unable to identify the optimal combination of superpixels. Similarly, if the size of the population is small, the algorithm may converge to a local optimum. If the size of the population is very large, the search area will be large, and the computational load will be high. To consider the effects of other parameters on the performance of genetic algorithms, we refer readers to [45] for a comprehensive introduction. The parameters in this paper were set on the basis of an empirical study. Table IV summarizes the value for each parameter.

TABLE IV
PARAMETERS SETTING

Parameter	Value
The size of the population N	20
The probability of crossover p_c	0.9
The probability of mutation p_m	0.09
The number of locations for mutation M	5
The number of super-pixels S	300
The maximum number of iteration T	50

A. Qualitative Analysis

Fig. 5 shows some examples to demonstrate the discriminative regions obtained by the proposed method, where the images in the first row are original fundus images, the images in the second row are images with superpixels, the images in the third row show the initial regions, and the images in the fourth row show the regions with the best fitness values. The images in the first two columns in Fig. 5 fall within the DR category, those in the third to fourth columns are examples that fall within the glaucoma category, and the those in the last two columns are normal examples. The proposed method is used to identify some superpixels with lesions, such as hard exudates and soft exudates, which are used as evidence for the diagnosis of DR in clinical practice, as discriminative regions. For glaucoma images, the proposed method is used to identify some superpixels located in optic disks as discriminative regions. In clinical practice, the ratio of the size of the optic disk to the size of the optic cup is commonly used for glaucoma diagnosis. Note that some superpixels that do not contain evidence used for making a diagnosis in clinical practice may be identified as discriminative regions by the proposed method. For example, the discriminative regions obtained with the proposed method for DR images are distributed among different locations rather than solely focusing on the regions with lesions.

Fig. 6 shows the performance curves for the examples given in Fig. 5. The horizontal axis in Fig. 6(a) represents the number of iterations, and the vertical axis represents the fitness value of the best individual in each iteration. The horizontal axis in Fig. 6(b) represents the number of iterations, and the vertical axis represents the pixel ratio for the best individual in each iteration. The pixel ratio is defined as the ratio between the number of preserved pixels (or discriminative pixels) in $mask$ obtained via the proposed method and the total number of pixels in an image. The fitness value improves as the number of iterations increases and tends to stabilize when the number of iterations is greater than a certain value. The curves for the pixel ratios exhibit fluctuations. This phenomenon occurs because crossover and mutation are used to generate new individuals, and the individual with the best fitness value in the current generation may differ from the one with the same best fitness value in the previous generation. In other words, different individuals may possess the same best fitness value. Fig. 7 shows four examples, which are the discriminative regions for the DR image shown in the first column in Fig. 5. These four discriminative regions are associated with different combinations of superpixels, even though they share the same best fitness value. This phenomenon indicates that the discriminative regions that can increase the confidence of CNNs for a specific category obtained via the proposed method are not unique.

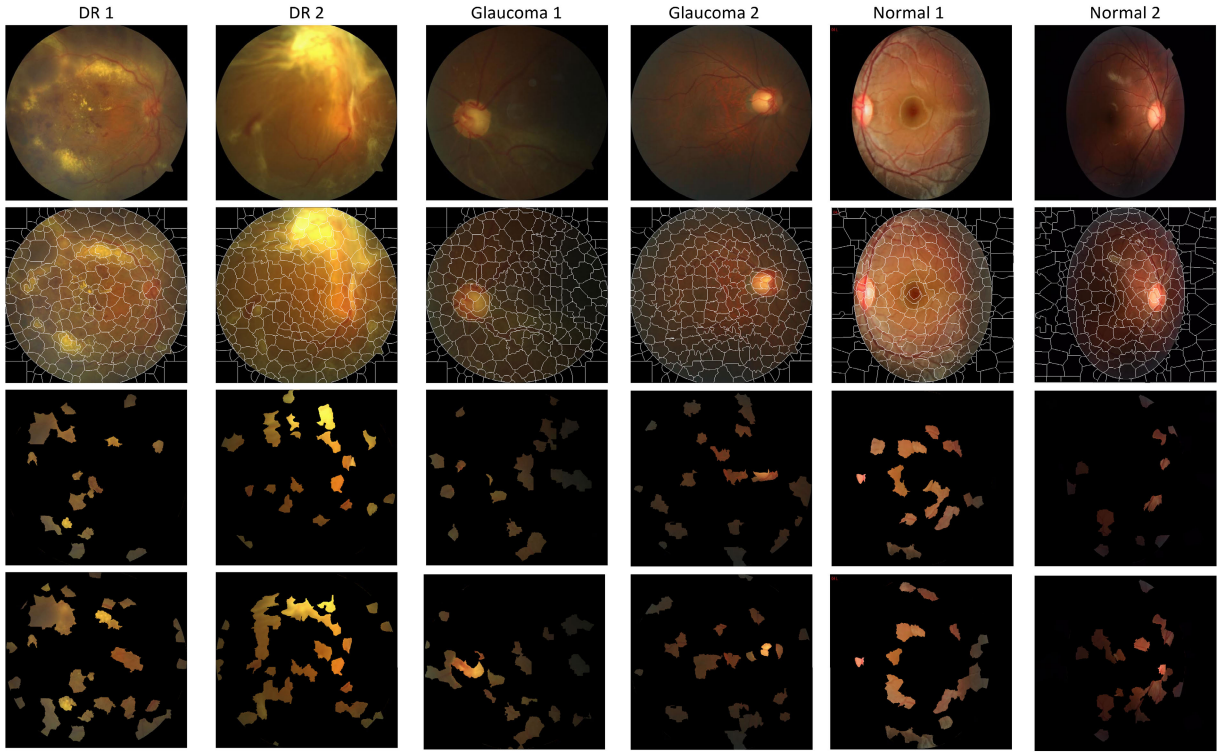


Fig. 5. Examples to demonstrate the discriminative regions obtained with the proposed method. The images in the first row are original images. The images in the second row are the images with superpixels. The images in the third row are the initial regions, and the images in the fourth row are the regions with the best fitness values.

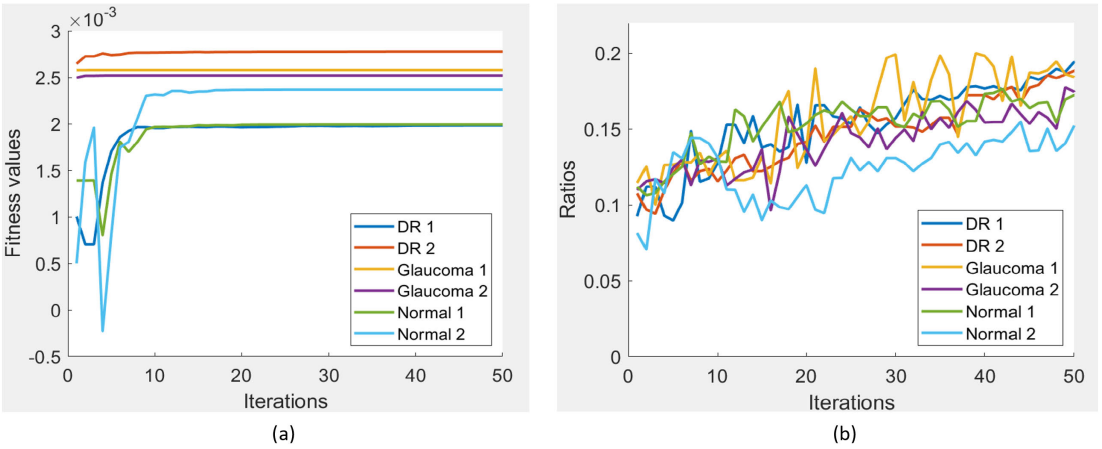


Fig. 6. The performance curves for each example in Fig. 5. (a) The horizontal axis represents the number of iterations. The vertical axis represents the fitness value of the best individual in each iteration. (b) The horizontal axis represents the number of iterations. The vertical axis represents the pixel ratio for the best individual in each iteration.

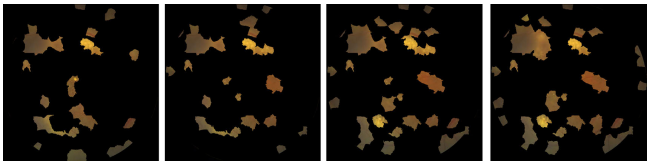


Fig. 7. Different discriminative regions with the same best fitness value for the DR image shown in the first column in Fig.5.

Notably, the pixel ratio for each image is influenced by the threshold value in the initialization step. Specifically, if the threshold value is small (resulting in more superpixels being preserved during initialization), the pixel ratio will be large. Conversely, if the threshold value is large (resulting in fewer superpixels being preserved during initialization), the pixel ratio will be small. Fig. 8 shows an example, where the curves are the performance curves for the DR image shown in the first column in Fig. 5 when the threshold value is set to 0.9, 0.5 and 0.1. A large threshold value results in a small pixel ratio. However, regardless of the threshold value, the best fitness values for the individuals tend to be the same.

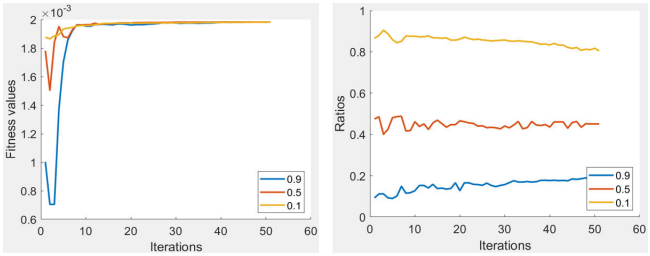


Fig. 8. The effect of the threshold value on performance. (a) Fitness values. (b) Pixel ratios.

TABLE V
THE PERFORMANCE OF THE PROPOSED METHOD IN SEARCHES FOR DISCRIMINATIVE REGIONS

Average Drop	Average Increase	Pixel Ratio
0	0.7780	0.1265

B. Quantitative Analysis

The average drop and average increase metrics used in [18] were employed to assess the quality of the proposed method. The average drop is defined as $\frac{1}{K}\sum_{k=1}^{K}\frac{max(0,Y_k^c-O_k^c)}{Y_k^c}$. The average increase (also defined as an increase in confidence) is defined as $\sum_{k=1}^{K}\frac{\mathbb{I}(Y_k^c<O_k^c)}{K}$, where Y_k^c is the predicted score for class c in image k and O_k^c is the predicted score for class c with the discriminative regions as inputs. $\mathbb{I}$ represents an indicator function that returns a value of 1 if the input is true and 0 otherwise. K represents the number of images in the test set. We also employed the average pixel ratios from the test set to assess the quality of the proposed method. Table V summarizes the results obtained with the proposed method in searches for discriminative regions of the CNN in fundus image classification. The average drop obtained with the proposed method is 0, which indicates that the proposed method can identify a discriminative region that does not decrease the confidence of the CNN for a specific category of feature for all the images in the test set. The pixel ratio obtained with the proposed method is 0.1265, which indicates that the CNN model can utilize a small portion of the pixels in an image to increase the confidence of classification for features in a specific category.

C. Comparison With Other Methods

Since most existing methods for explaining CNNs were evaluated with the ImageNet (ILSVRC2012) database [46], we randomly selected 200 images from the validation set in ImageNet to perform comparisons with other methods. The CNN model applied in this experiment was VGG-F, which was trained and can be downloaded from.[1] Note that the maximum number of iterations T was set to 1000 in this experiment to ensure that the algorithm tends to be as stable as possible for different images. Fig. 9 shows some examples to demonstrate the discriminative regions obtained with the proposed method applied to the ImageNet database, where the images in the first column are original images, the images in the second column are images with superpixels, the images in the third column show the initial regions, and the images in the fourth column show the regions with the best fitness values. The discriminative regions obtained with the proposed method are distributed in different locations in the images, which confirms the conclusion that CNNs employ different regional information from the images to increase the confidence for a specific category rather than solely focusing on the regions associated with a particular instance. Fig. 10 shows the performance curves for the examples given in Fig. 9. Similar to the cases shown in Fig. 6 for fundus images, the fitness values for natural images also improve and tend to stabilize as the number of iterations increases. In addition, the number of discriminative pixels obtained via the proposed method is relatively small, generally accounting for less than 25.

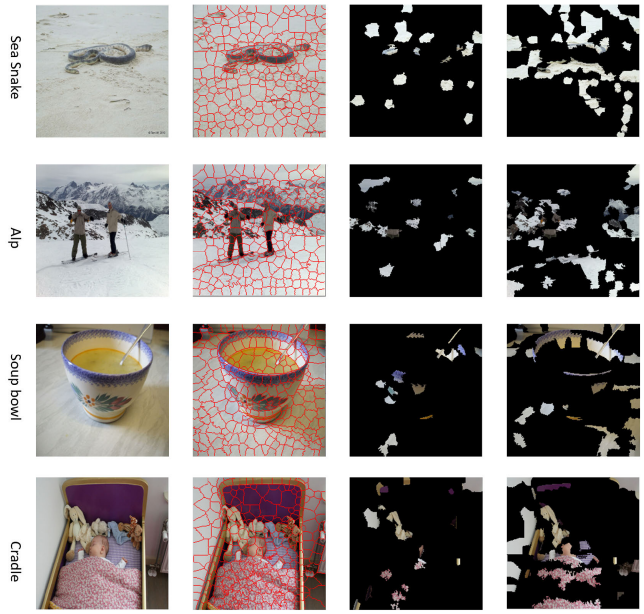
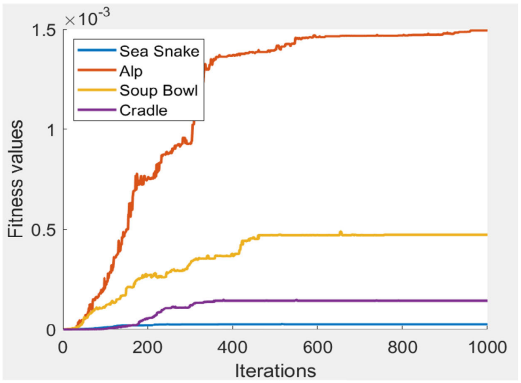


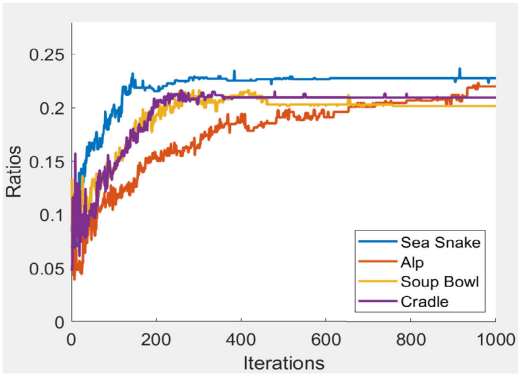
Fig. 9. Examples to demonstrate the discriminative regions obtained via the proposed method. The images in the first column are the original images. The images in the second column are the images with superpixels. The images in the third column show the initial regions. The images in the fourth column show the regions with the best fitness values.

Table VI summarizes the results obtained via different methods. In Table VI, the average drop and average increase for GradCAM [17], GradCAM++ [18], ScoreCAM [19], Rise [28] and Mask [14] are from [19]. Since the average drop and average increase were not provided in [27], we employed VGG-F and applied MO-LIME [27] to the same images input into the proposed model to calculate the average drop and average increase. In MO-LIME, a set of nondominated individuals is returned for each image, and the individual associated with the maximum increase in confidence for a specific category after 1000 iterations is selected for comparison. For the given metrics in Table VI, a lower average drop is better, and a higher average increase is better. The average drop and average increase obtained for the proposed method are significantly better than those obtained for other methods, such as CAM-based methods [17], [18], [19] and perturbation-based methods [14], [27], [28].

[1] https://www.vlfeat.org/matconvnet/



(a)



(b)

Fig. 10. The performance curves for the examples shown in Fig. 9. (a) The horizontal axis represents the number of iterations. The vertical axis represents the fitness value of the best individual in each iteration. (b) The horizontal axis represents the number of iterations. The vertical axis represents the pixel ratio for the best individual in each iteration.

TABLE VI
COMPARISON WITH DIFFERENT METHODS

Methods	Average Drop (%)	Average Increase (%)
GradCAM [17]	47.8	19.6
GradCAM++ [18]	45.5	18.9
ScoreCAM [19]	31.5	30.6
Rise [28]	47.0	14.0
Mask [14]	63.5	5.29
MO-LIME [27]	16.4	35.5
Proposed	2.97	84.5

As mentioned in Table I, the methods summarized in Table VI are all post hoc, local and model agnostic. The methods used for generating the discriminative regions differ among different methods. In the proposed method, the problem of searching for discriminative regions is formulated as a combinatorial optimization problem and is solved via evolutionary algorithms. The average drop and average increase obtained for the proposed method are significantly better than those for other methods, indicating that employing evolutionary algorithms to solve the combinatorial optimization problem is very effective in searching for discriminative regions that can increase the confidence of CNNs for classifying features in a specific category. In addition, although both MO-LIME [27] and the proposed method employ evolutionary algorithms to evolve discriminative regions, there are differences between these methods, such as the number of objectives, the methods used to achieve the objectives and the forms of the individuals. The average drop and average increase obtained for the proposed method are also significantly better than those obtained for MO-LIME, which indicates that the changes to the proposed method yield a significant improvement in the search for discriminative regions. Note that the computational cost of the proposed method primarily arises from the evaluation of individuals. When VGG-F is used as the analysis model, it takes approximately 26 s to perform 50 iterations with a population size of 20.

IV. DISCUSSION AND CONCLUSION

In this paper, we propose a method to search for discriminative regions and increase the confidence of CNNs for the classification of features in a specific category via genetic algorithms. In the proposed method, a set of superpixels is first generated for an image. Individuals are then initialized randomly for the superpixels. Genetic algorithms are employed to evolve individuals such that discriminative regions can be automatically identified. Many experiments are performed, and the results verify that the proposed method is highly effective in searching for discriminative regions that can increase the confidence of CNNs for the classification of features in a specific category.

In the proposed method, the discriminative regions identified are distributed in different locations in an image rather than focusing on regions associated with a particular instance, which may result in the causal relationship between the discriminative regions and the output of CNNs not being apparent. To enhance the clarity of this relationship, common features can be extracted from the discriminative regions for images classified in the same category by CNNs on the basis of the proposed method. Note that there are common features among the images classified in the same category by humans. For example, in the case of fundus images classified as DR, a common feature is the presence of related lesions.

CNNs may also identify unique features (which may not necessarily align with those used by humans) for images classified in the same category. The findings presented in this paper, including the observation that the discriminative regions for the same image are not unique and the relatively small pixel ratio for each image, lend some support to this supposition. How can we extract common features for CNN classification via the proposed method to better explain the behaviours of CNNs? This question issue will be explored in our future work.

REFERENCES

[1] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, J. H. Tan, and U. R. Acharya, "Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images," *Inf. Sci.*, vol. 441, pp. 41–49, May 2018.

[2] Z. Fan et al., "Optic disk detection in fundus image based on structured learning," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 224–234, Jan. 2018.

[3] Y. Hagiwara et al., "Computer-aided diagnosis of glaucoma using fundus images: A review," *Comput. Methods Programs Biomed.*, vol. 165, pp. 1–12, Oct. 2018.

[4] J. Kaur, D. Mittal, and R. Singla, "Diabetic retinopathy diagnosis through computer-aided fundus image analysis: A review," *Arch. Comput. Methods Eng.*, vol. 29, no. 3, pp. 1673–1711, May 2022.

[5] S. S. Rahim, V. Palade, C. Jayne, A. Holzinger, and J. Shuttleworth, "Detection of diabetic retinopathy and maculopathy in eye fundus images using fuzzy image processing," in *Proc. Int. Conf. Brain Informat. Health*, 2015, pp. 379–388.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–20.

[8] D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.

[9] L.-P. Cen et al., "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature Commun.*, vol. 12, no. 1, p. 4828, Aug. 2021.

[10] M. Gaur, K. Faldu, and A. Sheth, "Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?" *IEEE Internet Comput.*, vol. 25, no. 1, pp. 51–59, Jan. 2021.

[11] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[12] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Aug. 2020.

[13] T. Huber, K. Weitz, E. André, and O. Amir, "Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps," *Artif. Intell.*, vol. 301, Dec. 2021, Art. no. 103571.

[14] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3429–3437.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[16] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognit. Lett.*, vol. 150, pp. 228–234, Oct. 2021.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[18] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[19] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 24–25.

[20] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 972–980.

[21] M. Bany Muhammad and M. Yeasin, "Eigen-CAM: Visual explanations for deep convolutional neural networks," *Social Netw. Comput. Sci.*, vol. 2, no. 1, pp. 1–14, Feb. 2021.

[22] J. R. Lee, S. Kim, I. Park, T. Eo, and D. Hwang, "Relevance-CAM: Your model already knows where to look," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14944–14953.

[23] Q. Zheng, Z. Wang, J. Zhou, and J. Lu, "Shap-CAM: Visual explanations for convolutional neural networks based on Shapley value," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, 2022, pp. 459–474.

[24] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," *IEEE Signal Process. Mag.*, vol. 39, no. 4, pp. 73–84, Jul. 2022.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[26] M. T. Ribeiro, S. Singh, and G. Carlos, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–20.

[27] B. Wang, W. Pei, B. Xue, and M. Zhang, "A multi-objective genetic algorithm to evolving local interpretable model-agnostic explanations for deep neural networks in image classification," *IEEE Trans. Evol. Comput.*, vol. 2, no. 1, pp. 1–15, May 2022.

[28] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[29] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[30] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," *arXiv preprint arXiv:1807.08024*, 2018.

[31] J. D. Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, and A. Holzinger, "On generating trustworthy counterfactual explanations," *Inf. Sci.*, vol. 655, Jan. 2024, Art. no. 119898.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[33] C.-C.-J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 406–413, Nov. 2016.

[34] W. L. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 2217–2225.

[35] Z. J. Wang et al., "CNN explainer: Learning convolutional neural networks with interactive visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1396–1406, Feb. 2021.

[36] X. Xuan, X. Zhang, O.-H. Kwon, and K.-L. Ma, "VAC-CNN: A visual analytics system for comparative studies of deep convolutional neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 6, pp. 2326–2337, Jun. 2022.

[37] O. Deperlioglu, U. Kose, D. Gupta, A. Khanna, F. Giampaolo, and G. Fortino, "Explainable framework for glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation," *Future Gener. Comput. Syst.*, vol. 129, pp. 152–169, Apr. 2022.

[38] J. Chang et al., "Explaining the rationale of deep learning glaucoma decisions with adversarial examples," *Ophthalmology*, vol. 128, no. 1, pp. 78–88, Jan. 2021.

[39] M. S. Kamal, N. Dey, L. Chowdhury, S. I. Hasan, and K. Santosh, "Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[40] R. Ibrahim and M. O. Shafiq, "Explainable convolutional neural networks: A taxonomy, review, and future directions," *ACM Comput. Surveys*, vol. 55, no. 10, pp. 1–37, Oct. 2023.

[41] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[42] J. H. Holland, "Genetic algorithms," *Sci. Amer.*, vol. 267, no. 1, pp. 66–73, 1992.

[43] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Jul. 2012.

[44] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 689–692.

[45] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. B. S. Prasath, "Choosing mutation and crossover ratios for genetic algorithms—A review with a new dynamic approach," *Information*, vol. 10, no. 12, p. 390, Dec. 2019.

[46] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.