IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 10, 2020, accepted March 1, 2020, date of publication March 13, 2020, date of current version April 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980742

Salient Explanation for Fine-Grained Classification

KANGHAN OH[1], SUNGCHAN KIM[1], AND IL-SEOK OH[1,2]

[1]Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[2]Research Center for Artificial Intelligence Technology, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding author: Il-Seok Oh (isoh@jbnu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) through the Korean Government, Ministry of Science and ICT (MSIT) under Grant 2019R1F1A1063522.

ABSTRACT Explaining the prediction of deep models has gained increasing attention to increase its applicability, even spreading it to life-affecting decisions. However there has been no attempt to pinpoint only the most discriminative features contributing specifically to separating different classes in a fine-grained classification task. This paper introduces a novel notion of salient explanation and proposes a simple yet effective salient explanation method called Gaussian light and shadow (GLAS), which estimates the spatial impact of deep models by the feature perturbation inspired by light and shadow in nature. GLAS provides a useful coarse-to-fine control benefiting from scalability of Gaussian mask. We also devised the ability to identify multiple instances through recursive GLAS. We prove the effectiveness of GLAS for fine-grained classification using the fine-grained classification dataset. To show the general applicability, we also illustrate that GLAS has state-of-the-art performance at high speed (about 0.5 sec per 224 × 224 image) via the ImageNet Large Scale Visual Recognition Challenge.

INDEX TERMS Computer vision, neural networks, explainable artificial intelligence, machine learning.

I. INTRODUCTION

Over the last several years, convolutional neural networks (CNNs) [1] have achieved superior performance in various computer vision tasks, including image classification [2], [3], object detection [4], and image captioning [4]. Despite these dramatic advances, the opacity of CNNs makes it difficult to understand why they reach particular decisions, limiting the ability to widen their application to various fields.

In general, the visual interpretation of deep learning models is understood as estimating the impact of a particular neuron activation related to a given input instance. In white-box approach, architectural modification of the classification model [6] or access to specific layers [6], [9], [13] is inevitable [14], resulting in severe limitation of application. In contrast, the black-box approach [14]–[19] aims to be inherently model agnostic. Its main concerns are how to perturb an input image and draw the model's response on the perturbed instance to the final heat map. For example, the Randomized Input Sampling for Explanation (RISE) method [14] perturbed an image with a randomised mask to measure the importance of pixels and then linearly fused all importance from several thousand masks.

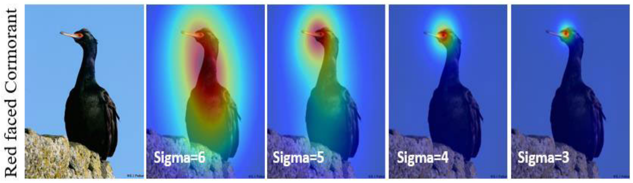

FIGURE 1. Salient explanation. As the value of σ decreases, the heat map concentrates more and more on the most discriminative part of the relevant objects, the bird's face in this case. The salient explanation is very important for a variety of tasks such as the fine-grained classification and biomarker discovery in medical image. Note that the salient explanation is possible due to our idea of adopting Gaussian mask.

The conventional black-box methods employed unnatural and fragile perturbation schemes such as single colour out [14], [16], [17], random noise [18], [19], and smoothing [19]. These perturbation schemes have several limitations. First, they are deficient in pinpointing only the most discriminative, i.e., salient features that are essential for the fine-grained classification tasks where the between-class shape similarity is very high — for example, pinpointing only the red face of Red-faced Cormorant in the bird classification task in Fig. 1 is crucial for explaining why a deep learning model classifies

The associate editor coordinating the review of this manuscript and approving it for publication was Syed Muhammad Anwar.

the image as Red-faced Cormorant. Second, the conventional perturbation schemes highly suffer from local noise and, thus, fuse maps from a considerable number of perturbations for a reliable explanation. This is the main cause of slowness with conventional black-box methods.

Inspired by the lighting and shadowing phenomena in nature, we propose a simple yet effective black-box method, called Gaussian light and shadow (GLAS), which simulates feature perturbation as the presence or the absence of light at the pixel level of an image. The primary idea of GLAS is to perturb an input image by the Gaussian mask (light) and inverse-Gaussian mask (shadow) and, then, record the responses of the perturbed images. GLAS uses a simple grid search; once completed over the entire image, the response maps are fused to construct the final heat map. The fusion mimics the Gaussian mixture. The proposed method has several advantages compared with other black-box methods [14], [18], [19]. First, GLAS provides scalability of explanation that we can achieve by adjusting the variance parameter of the Gaussian mask. The scalability makes it possible to pinpoint clues for the salient explanation, which is not feasible with the nonparameterized approaches [14], [17]–[19]. The salient explanation is valid for explaining fine-grained classifications, such as classifying bird species in a CUB200 dataset, involving large between-class similarity and significant within-class variance. Fig. 1 shows an image of the Red-faced cormorant species that we can discriminate by identifying the face color. It illustrates that GLAS adjusts its gaze from the body to the red face as the scale parameter decreases and finally pinpoints the red area around the eye. Second, our pixel-wise multiplication operation with the Gaussian mask at a specific search point simulates the gradual dimming effect as going farther from the center. We argue that because of this characteristic, a significantly reduced number of perturbations is sufficient. GLAS can process an image much faster than conventional methods.

To summarize, the contributions of this paper are as follows:

1) We introduce a novel notion of salient explanation which is critical in explaining the fine-grained classification tasks. We propose a simple yet efficient black-box method, GLAS, which provides an easy way to perturb an input image based on Gaussian lighting and shadowing.
2) GLAS is fast because of the smoothly varying shape of the Gaussian mask, which generates a visual explanation up to one order of magnitude faster than other black-box methods.
3) We show the broad applicability of GLAS to various other tasks: object localization and visual captioning. Quantitative comparisons show that GRAS is superior to conventional methods.

II. RELATED WORKS

The white-box approach heavily uses the network's internal information, such as gradients or feature maps of specific layers. A gradient can indicate how much a small change in a pixel influences the class output [20]. For example, Simonyan and Zisserman [10] proposed the gradient-based model, which directly mapped saliency values to the original space. Additionally, Zeiler and Fergus proposed a deconvolution method [17]. In the method, the forward signal is reversed at a neuron and backpropagated to the input space. The study [6] proposed the layer-wise relevance propagation method, in which the prediction in the output layer is decomposed into pixel-wise relevance values and backpropagated until it satisfies the conservation rule. Samek *et al.* [21] emphasized the importance of quantitative evaluation and provided a rigorous comparison of the previously mentioned methods; these approaches are extensively reviewed in Samek *et al.* [21]. The visual feature maps provide important clues for the explanation, which some techniques exploit. The technique [4] called class activation mapping (CAM) is accomplished by weighted fusion of visual feature maps and requires the modification of CNN architecture, replacing the fully connected layer with the global average pooling. Grad-CAM [9], an extended version of CAM, is applicable to a broader range of CNNs. The previously mentioned techniques modify the model's internal operations or rely on the model's internal values; thus, they are model dependent.

The black-box approach measures the response change of the base model when the input instance is spatially perturbed, and this change can be regarded as the significance of the classifier's decision. The study [12] simulates feature perturbation based on marginal probability, and several studies have extended and improved this method [18], [22]. For the CNN-based architectures, the study [18] proposed the conditional sampling approach. The method considers that a given pixel value highly depends on neighboring pixels and that multivariate analysis excludes a rectangular region rather than a single pixel. Because a pixel-wise perturbation method such as random noise [18], [19], was considered, pixels are highly vulnerable to adversarial attack.

Several techniques [15], [16], [22], aim at region-wise perturbation approaches, rather than using pixels. A study [22] improved the conditional sampling method using the superpixel algorithm, making it more robust from local noise than Zintgraf's method [18]. Additionally, the superpixel segmentation technique was used in existing methods [15], [16], [22]. In these methods, high-level segments, rather than pixels [18] or oversegmented regions [22], are used to perturb the feature of instance, and the methods have achieved explanation results that are more visually pleasing compared with previous methods. However, the results are probably limited when the segmentation map's quality is poor. Petisiuk *et al.* proposed the RISE method [14], which simulates the feature's absence using randomized masks and measures its response to each masked instance. Because of its random masking strategy, RISE requires a considerable number of feed-forward executions and suffers from local noise. The meta learning approach that tries to maximize the interpretability of a learning model is used in some studies

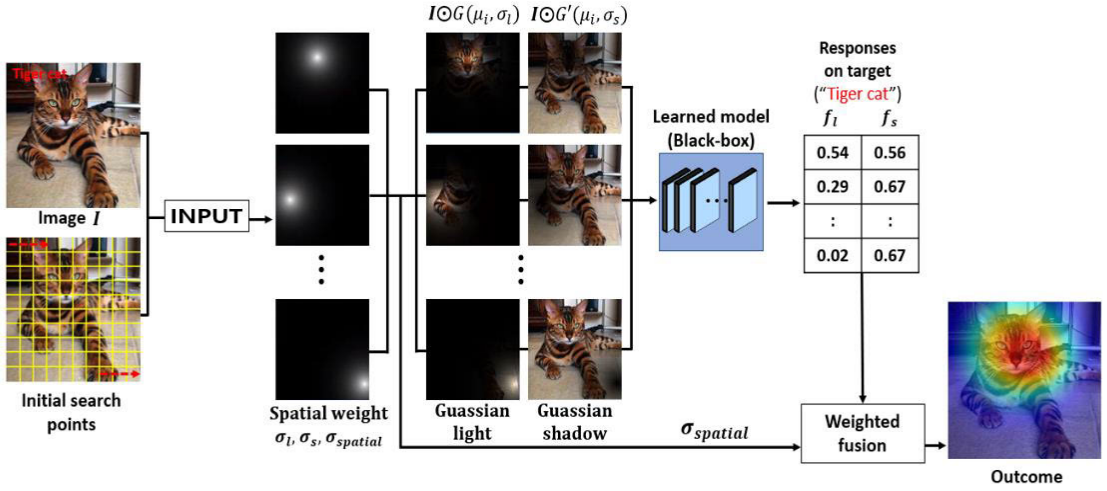

FIGURE 2. **Overview of the GLAS method.**

[16], [19]. One study [16] employed superpixel-wise random samples around the instance and an approximate linear decision model. Fong and Andrea [19] proposed an optimized framework that learns a minimum perturbation mask from the corresponding response to its output neuron. However, such frameworks often fail to optimize their result because of its sensitivity to various types of models and instances. Unlike the white-box approaches, the black-box methods are inherently model agnostic; i.e., they are applicable to any learning model, because they rely only on the output values, regardless of the internal workings of the classification models.

The fine-grained classification is to recognize subordinate classes of a base class such as species of birds and different models of cars and planes. Most of recent works use the deep CNN models and propose better loss function. Shi *et al.* [3] proposed a generalized large-margin (GLM) loss to reduce between-class similarity and within-class variance. The contrastive loss [23] and triplet loss [24] have also been proposed. Qiu *et al.* [25] proposed a method based the sqeeze-and-excitation attention model. Peng *et al.* [26] used both the object-level attention and part-level attention. The literatures treated various types of objects. Shi *et al.* [3] used birds, cars, and airplanes datasets. Other objects include fish [25], vehicle [27], plant [28], leukemia [29], and plankton [30].

III. PROPOSED APPROACH

A. GLAS METHOD

Given an image $\boldsymbol{I}$, we define a set of search points $M = (\mu_1, \mu_2, \ldots, \mu_{k\times k})$ by the centers of $k\times k$ grids overlaid upon $\boldsymbol{I}$, as shown in Fig. 2. For a given class label y and a specific search point μ_i, the prediction score $f_l(\mu_i)$ by Gaussian light can be written as

$$f_l(\mu_i) = P\left(y|\mathbf{I}\odot G(\mu_i,\sigma_l)\right) \tag{1}$$

where $\odot$ denotes element-wise multiplication, and $G(\mu_i, \sigma_l)$ is a Gaussian distribution with mean μ_i and standard deviation σ_l. Equation (1) simulates light projected on a specific part of the image to measure the contribution of the local pattern. It is also possible to define the score based on the inverse-Gaussian mask; i.e., the shadow is given by the following equation:

$$f_s\left(\mu_i\right) = \left|P\left(y|\mathbf{I}\right) - P\left(y|\mathbf{I}\odot G'(\mu_i,\sigma_s)\right)\right| + \gamma \tag{2}$$

where $G^{'}\left(\mu_i,\sigma_s\right) = 1 - G(\mu_i,\sigma_s)$. Here, γ is a constant, and we empirically set it to 10^{-5} to avoid $f_s\left(\mu_i\right)$ being 0. We use a weighted fusion to define the saliency score $S(x_j)$ for a pixel x_j as

$$S(x_j) = \frac{1}{|M|}\sum_{\mu_i\in M}\exp\left(-\frac{D\left(x_j,\mu_i\right)}{\sigma^2_{spatial}}\right)f_l(\mu_i)f_s(\mu_i) \tag{3}$$

where $\exp\left(-\frac{D(x_j,\mu_i)}{\sigma^2_{spatial}}\right)$ is a spatial weighting factor; here, $D(a,b)$ denotes the distance between a and b. Equation (3) represents the Gaussian mixture-based weighted fusion. The high flexibility of the visual explanation can be achieved by adjusting the scale parameter σ for each Gaussian mask.

Algorithm 1 RGLAS

Input: $\boldsymbol{I}$, class label y

Output: Saliency map $\boldsymbol{S_{fuse}}$

```
$S_{fuse} = 0$; $M = (\mu_1, \mu_2, \ldots, \mu_{k\times k})$
Repeat
  for each pixel $x_j$ in $I$
     $S\left(x_j\right) = \frac{1}{|M|}\sum_{\mu_i\in M}\exp\left(-\frac{D(x_j,\mu_i)}{\sigma^2_{spatial}}\right)$
     $f_l(\mu_i) f_s(\mu_i)$
   $B = S > t_1$
   for each search point $\mu_i$ in $M$
      If $(B\left(\mu_i\right) == 1)$ remove $\mu_i$ from $M$
      $I = I\odot(1-S)$
     If $\left(\sqrt{\frac{mean(S)}{f(y|I)}} > t_2\right)$ break
     else $S_{fuse} += normalize\ S$
   Normalize $S_{fuse}$
```

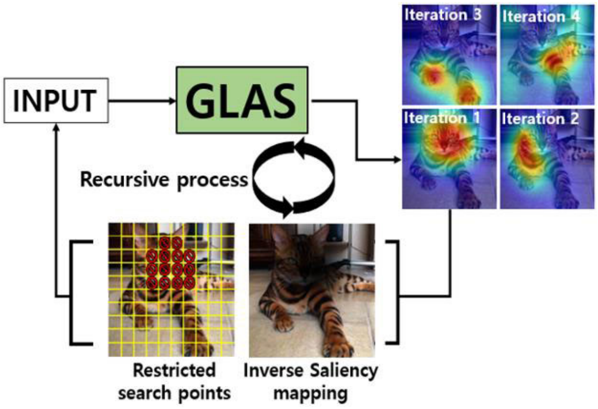

FIGURE 3. Framework of RGLAS. The GLAS instances are repeated until all discriminative patterns related to a given class have been discovered.

B. RECURSIVE GLAS METHOD

GLAS tends to highlight the most discriminative clue. To discover the various evidences that lead to the classifier's decision, we propose a simple schema called RGLAS. Fig. 3 shows the key idea of RGLAS: to prevent revisits to the search points related to the most discriminative features that have already been found, leading to extraction of the next important features. This mechanism also helps in discovering multiple instances in an image. The RGLAS algorithm starts by constructing the saliency map S (lines 3–4). We compute the binary map of S using the threshold value $t_1 = 0.8$ (line 5) and eliminate search points located in the positive region of the binary map (lines 6–7). The input image is updated using the previous input and the inverse saliency map (line 8). We define a simple stop condition, as formulated in line 9, with $t_2 = 5$. We found that as the iteration increases, mean(S) tends to increase but $f(y|I)$ decreases, guaranteeing that the stop condition occurs consistently.

IV. EXPERIMENTAL RESULTS

The experiments were conducted on an Intel Core i7-7800X with a 3.50 CPU, 32 GB of memory, and a GTX 1080 Ti GPU. We aimed to evaluate quantitatively and qualitatively the salient explanation capabilities of GLAS and existing explanation models.

A. SALIENT EXPLANATION FOR FINE-GRAINED CLASSIFICATION TASKS

The GLAS provides us with fine-level visual clue identification, enabling the salient explanation. To demonstrate the effectiveness of scalability, we employed CUB200 [31], Stanford Cars [32], and Aircraft [33] benchmarks that have been used for the fine-grained image classification tasks. The CUB200 dataset consists of 11,788 images of 200 bird species. The Stanford Cars dataset includes 8,144 training and 8,041 test images with 196 classes. The Aircraft dataset is a set of 10,000 images with 100 classes reflecting a fine-grained set of airplanes. We used the basic procedure of transfer learning using ResNet-50 pretrained on ILSVRC. The resulting networks of CUB200, Stanford Cars and Aircraft yielded top-1 accuracies of 76.11%, 92.01%, and 80.59%, respectively. GLAS can pinpoint more detailed clues by adjusting its scale parameters. Fig. 4 shows the visual comparison. In the experiment, we used an equal sigma value of 3.0 for σ_l, σ_s, and $\sigma_{spatial}$. Unlike other methods, we can see that GLAS consistently pinpoints meaningful patterns of birds. For example, in the case of the Northern flicker, an instance has characteristics such as the red spot below the eyes and black dots on the body. GLAS surprisingly pinpoints two characteristics of the Northern flicker with the scale 3.0; however, the other methods only discover the location of the instance and fail to explain the detailed patterns. In this regard, Grad-CAM and RISE tend to explain

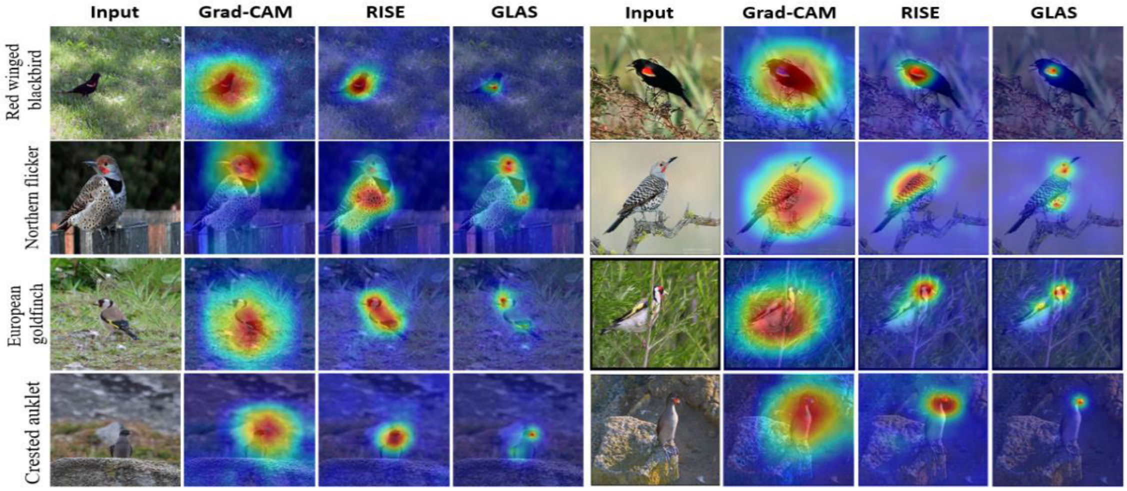

FIGURE 4. Visual comparison with the existing models. From top to bottom: red-winged blackbird, Northern flicker, European goldfinch, and crested auklet. We introduce the uniform characteristics of each bird: first row (red wings), second row (red below the eyes and black spots on the body), third row (red face and yellow on the wings), and fourth row (orange beak).

TABLE 1. Average distances between the landmark points of selected categories and the maximum points of heat maps.

Category	Landmark	Grad -CAM	RISE	GLAS (sigma 5)	GLAS (sigma 3)
Red-winged blackbird	Wings	41.14	19.34	18.97	16.25
Crested auklet	Beak	38.89	21.41	20.68	17.20
Red-faced cormorant	Head	39.77	15.02	16.01	13.37
European goldfinch	Head	51.62	39.77	42.33	33.31
Eared grebe	Eye	48.73	28.88	30.14	21.01

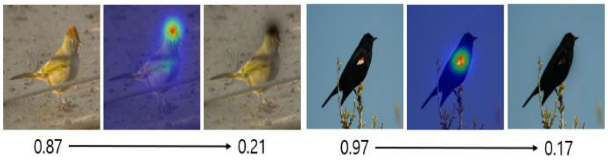

FIGURE 5. From left to right: original image, heat map, and image perturbed by the inverse heat map along with the class probabilities. Green-tailed towhee and red-winged blackbird.

only global significance, such as the target's location. Numerous visual examples are available in the Supplementary Materials.

The CUB200 dataset provides 15 landmark points for each bird: beak, crown, eyes, nape, etc. We employed these annotations to evaluate quantitively the salient explanation capability of GLAS. We chose five categories in Table 1 due to their unique characteristics, e.g., red-wing blackbird with red spot on the wing. The landmark of each bird corresponding to the unique characteristic is described in second column. We measured the Euclidian distance between the landmark point and the maximum point of the heat map. As Table 1 shows, GLAS with sigma = 3.0 ($\sigma_l = \sigma_s = \sigma_{spatial} = 3.0$) achieved the shortest distance compared with both other methods and GLAS with sigma = 5.0. The results tell us GLAS can closely access the meaningful patterns of birds. We conducted another experiment measuring how much the pinpointed clues affect the class decision. In Fig. 5, original images are perturbed by the inverse heat map, and the amount that the score drops is recorded. As expected, the class score dropped rapidly, ensuring the significance of the pinpointed features. Fig. 6 shows the European goldfinch characterized by red face and yellow spot on the wings used for demonstrating coarse-to-fine controls. GLAS adjusts the standard deviation. Deconv adjusts the occlusion mask size. The RISE adjusts the size of the initial mask. The most prominent observation is that the heat maps from RISE and Deconv are very noisy and less accurate in identifying the most discriminative parts of the relevant object. The failure case analysis in Fig. 7 deserves an attention. The second row unveils an interesting behavior of CNN through failure cases. The first and second images belonging to the Red-legged Kittiwake was incorrectly classified into Pigeon guillemot with 34.09% and 11.59% probabilities, respectively. The salient explanation capability of GLAS allows us to understand that the CNN misclassified the images into the Pigeon guillemot by looking at the red leg. Fig. 8 shows the visual explanations on Aircraft and Stanford Cars benchmarks, respectively. For Aircraft examples, GLAS consistently pinpointed propellers of wings for the class "Yu 12". For Stanford Cars dataset, we found that CNN changed its gaze consistently according to poses of Car. When the front is shown for the car class 138, the grille part is highly probable to be pinpointed. When the back side is shown, lamps and wheels are pinpointed. Note that these behaviors of CNNs can be unveiled only when the salient explanation is available.

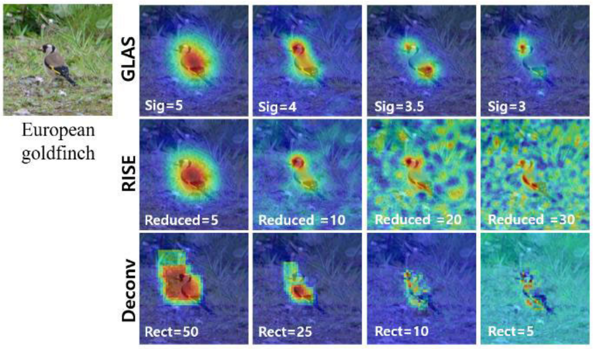

FIGURE 6. Visual comparisons of GLAS, RISE, and Deconv according to their parameters controlling the locality. The example is the European goldfinch characterized by red face and yellow spot on the wings. GLAS adjusts the standard deviation. Deconv adjusts the occlusion mask size. The RISE adjusts the size of the initial mask.

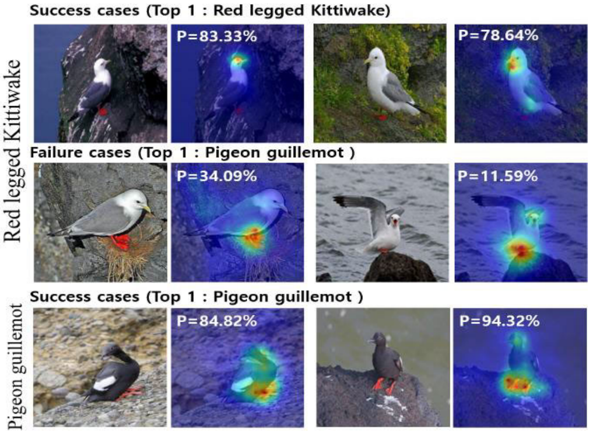
Success cases (Top 1 : Red legged Kittiwake)
P=83.33%
P=78.64%
Failure cases (Top 1 : Pigeon guillemot)
P=34.09%
P=11.59%
Success cases (Top 1 : Pigeon guillemot)
P=84.82%
P=94.32%
Red legged Kittiwake
Pigeon guillemot

FIGURE 7. Unveiling the feature selection behaviors of CNN. The top two rows illustrate two bird classes, Red-legged Kittiwake and Pigeon guillemot. Despite its name, the Red-legged Kittiwake were consistently highlighted on faces while the Pigeon guillemot was highlighted on the red legs.

B. EVALUATION ON TARGET LOCALIZATION

We performed quantitative evaluations on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). As a metric, we employed the pointing game (PT) presented in the study [13]. The PT purely measures the spatial selectiveness of the continuous visual saliency map. In the evaluation, the PT detects the maximum intensity point on the saliency map, and a Hit is recorded if the maximum point is in the ground-truth annotation; otherwise, a Miss is recorded. The accuracy is calculated using $Acc = \frac{\#Hit}{\#Hit+\#Miss}$.

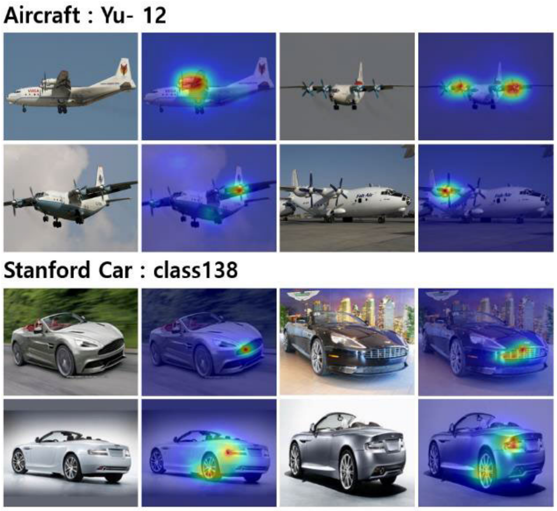

FIGURE 8. Visual explanations on Aircraft and Stanford Car benchmarks.

TABLE 2. PT scores according to the number of search points.

Search points $k \times k$	12×12	15×15	22×22	25×25	30×30
PT score	0.905	0.912	0.914	0.913	0.913
Time (s)	0.398	0.652	1.435	1.909	3.951

TABLE 3. Quantitative comparisons with existing models using the PT on the ILSVRC validation data using ResNet50.

Method\Size n	PT	PT-small	Time (s)
Grads [10]	0.773	0.604	0.128
Deconv [17]	0.750	0.584	0.117
Grad-CAM [9]	0.901	0.754	0.183
Deconv [17]	0.809	0.688	2.64
LIME [16]	0.766	0.645	15.11
RISE [14]	0.907	0.787	8.11
MASK [19]	0.841	0.711	16.38
GLAS (light)	0.889	0.782	0.328
GLAS (shadow)	0.877	0.751	0.328
GLAS (fusion)	**0.912**	0.792	0.612

Because multiple maximum points often arise, we employed the threshold value $T > 0.95$ to generate binary blobs, and then we used the centroid of the biggest blob as the localization point.

We empirically set the scale parameters $\sigma_l = 5$ and $\sigma_s = 3$, with $\sigma_{spatial} = 6$. Table 2 shows the execution time of GLAS according to the number of grid search points. The result tends to show favorable PT scores as the number of the search points grows. It starts to become saturated after $k = 15$ in terms of performance, even as the execution time continues to grow. Table 3 illustrates the results of comparisons with the existing methods on the ILSVRC validation dataset. GLAS achieves the best result in terms of PT score. GLAS is 13 times faster than RISE even with the higher PT score. This is because a considerable number of perturbed images using a randomized masking process are necessary for reliable visual explanation in RISE. When RISE is forced to use 450 (225×2) perturbed images, identical to the number used by GLAS, we observed that the PT score of RISE drops from 0.907 to 0.869. This observation tells us that GLAS perturbs and localizes the important features efficiently.

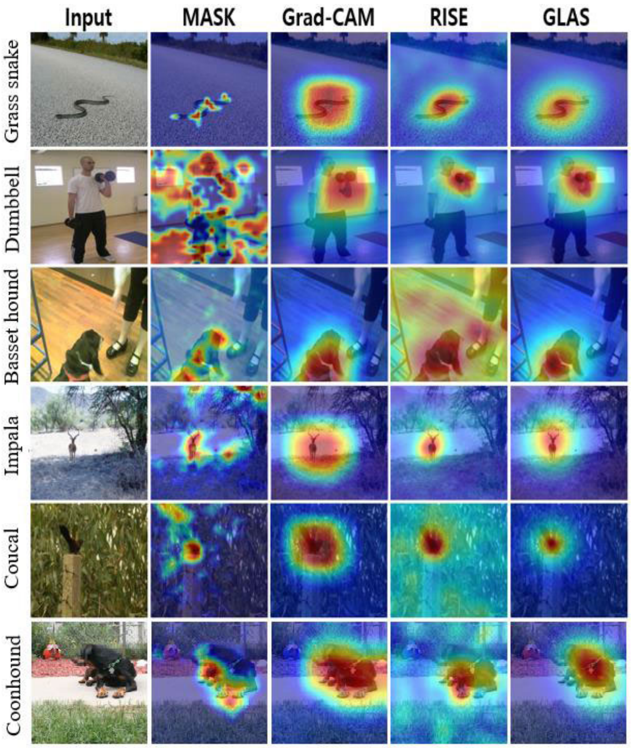

FIGURE 9. Visual comparison of the class-discriminative capability of MASK, Grad-CAM, RISE, and GLAS by varying the object classes.

We separately evaluated the cases in which the object is small in the PT-small column of Table 3. We consider an object to be small if the total area of the bounding box of the given class is smaller than one quarter of the size of the image. Even though all models encountered a performance drop, GLAS still beats other models. In our work, GLAS operations can be used together or independently. The results in the last three rows of Table 3 show an ablation study on GLAS by measuring the performance with either Gaussian lighting or shadowing suppressed. Table 3 shows the performance of the ablated GLAS is comparable to that the state-of-the-art methods, outperforming all methods except Grad-CAM, RISE, and fused GLAS.

Fig. 9 provides a visual comparison of the methods. GLAS and Grad-CAM clearly highlight the important region related to a given class, whereas MASK and RISE suffer from non-trivial local noise. The advantage of GLAS is obvious without the noise and produces a visualization map that is highly interpretable. Because the GLAS map consists of Gaussian mixture clues, it identifies the most important area without being distracted by meaningless clues.

C. MULTIPLE EVIDENCE DISCOVERY BY THE RECURSIVE PROCESS

Table 4 illustrates the mean intersection over union (IOU) scores of the proposed recursive process. Because the visual saliency map consists of continuous intensity values,

TABLE 4. IOU scores of the proposed method on the ILSVRC validation data as the iteration increases. "A" represents the final iteration under the adaptive stop condition.

Iteration	Threshold				
	0.1	0.3	0.5	0.7	0.9
1	0.53	0.51	0.40	0.31	0.09
2	0.54	0.56	0.51	0.39	0.07
3	0.55	0.56	0.52	0.37	0.06
A	0.55	0.59	0.55	0.38	0.06

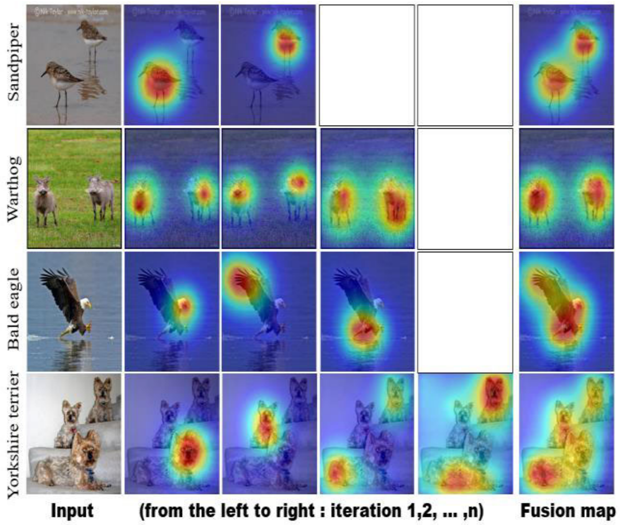

FIGURE 10. Recursive process. This algorithm discovers the relevant parts of a given class in order of significance. We used the adaptive stop condition mentioned in Section 3.

the mean IOU scores were measured with varying thresholds, from 0 to 1.0. In particular, significant progress occurs at the second iteration. Higher IOU scores over the threshold values indicate that the object regions are uniformly highlighted. In the fusion results shown in the last column of Fig. 10, we can see that multiple instances and multiple evidences are well discovered.

D. VISUALIZATION OF THE IMAGE CAPTIONING MODEL

Image captioning is a challenging task for which both computer vision and natural language processing techniques should be considered [34]. We constructed the image captioning model based on publicly available implementations[1] for which the fine-tuned InceptionV3-based image and long short-term memory-based language models are considered. Fig. 11 shows some visual explanation results from the image captioning model to demonstrate the applicability of GLAS. GLAS shows the capability to localize visual concepts such as objects (man, bicycle, ball, boy, hurdle, and dog) and actions (riding, playing, and jumping).

E. VISUALIZATION OF THE NEUROIMAGING CLASSIFICATION MODEL

We consider a neuroimaging classification problem to show the applicability of GLAS in a medical imaging domain. We employed 3D-magnetic resonance imaging (MRI) scans reflecting 199 patients with Alzheimer's disease (AD) versus 230 healthy normal control (NC) individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which is publicly available. A 3D-CNN was employed for classification of AD versus NC. Because of the limited dataset size, 3D-MRI scans are spatially normalized based on template brain image, and the unsupervised learning technique (convolutional autoencoder) is applied before supervised learning [35]. The overall architecture is composed of three 3 × 3 × 3 Conv layers with 10 filters each, two FC layers with 32 and 16 nodes each, and softmax activation. Fivefold cross-validations were conducted to evaluate the classification model. The mean accuracy was 85.24% [37].

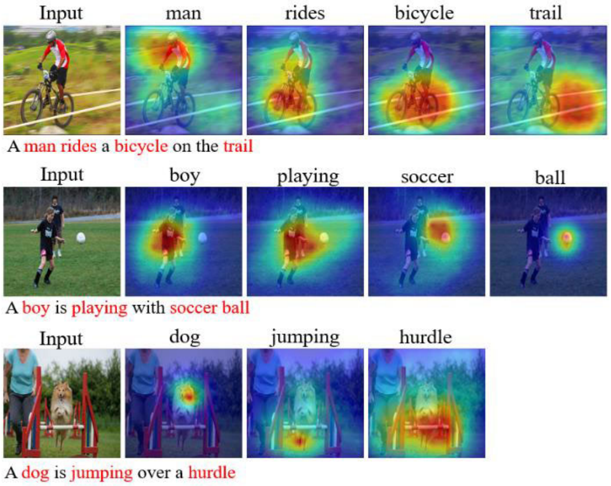
Input
man
rides
bicycle
trail
A man rides a bicycle on the trail
Input
boy
playing
soccer
ball
A boy is playing with soccer ball
Input
dog
jumping
hurdle
A dog is jumping over a hurdle

FIGURE 11. Visual explanation examples produced by GLAS for the image captioning model.

In applying GLAS to MRIs, we used the automated anatomical labeling (AAL) map which contains 116 anatomical segments representing brain functioning. We considered the centroid of each segment as the search point rather than use of grid point. The 3D-GLAS used 3D Gaussians to perturb MRI. We empirically set $\sigma_l = 10$ and $\sigma_s = 15$. Regarding a segment μ_i as unit, all the voxels in the segment were assigned the same saliency score, $f_l(\mu_i, \sigma_l)f_s(\mu_i, \sigma_s)$. For a statistical analysis of AD group, the MRI scans of the AD category are fed to 3D-GLAS, and the saliency maps were linearly combined and normalized. In the second row of Fig. 12, the segments corresponding to hippocampus, amygdala, and temporal inf were highlighted as the important biomarkers for the accurate classification of AD. The first row of Fig. 12 shows these biomarkers with different colors. These biomarkers have been proved to be closely related to dementia in many studies [36]–[38]. In particular, the hippocampus, a brain region for learning and memory, is one of the first brain biomarkers affected by AD, and it undergoes severe structural changes as the disease progresses. This experiment shows that GLAS is able to interpret the learned neuroimaging classification model based on AAL.

[1] https://github.com/yashk2810/Image-Captioning

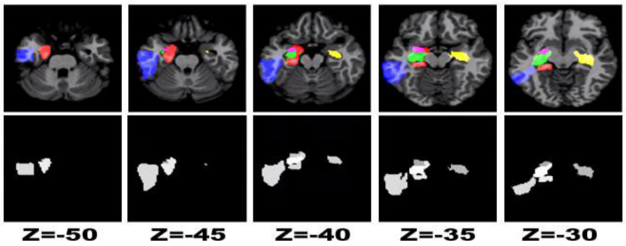

FIGURE 12. **Visual distribution of discriminative biomarkers in the classification of AD. The first rows illustrate a brain template image overlapped with important biomarkers (rank 1 of Table 1, red; rank 2, green; rank 3, blue; rank 4, yellow; rank 5, purple). The second rows show the AAL-wise saliency map. Z represents depth of MRI scan, Z=0 corresponding to the middle cross section, and $Z = -1, -2, -3, \ldots$ going to down.**

V. CONCLUSION

In this study, we proposed a visual explanation method called GLAS for the black-box model. Our method is inspired by the natural light and shadow phenomena and provides a simple yet robust way to perturb an input instance. In experiments, GLAS showed state-of-the-art performance and efficient computing time. In particular, the GLAS presented the ability of fine-level visual explanation at various scales through the adjustment of the Gaussian scale. Additionally, we showed the wide applicability of GLAS to various tasks. For a future work, we plan to improve the saliency map by optimizing the scale parameters of the Gaussian mask adaptively on an image instance. A deeper theoretical analysis of GLAS is also needed.

REFERENCES

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

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] W. Shi, Y. Gong, X. Tao, D. Cheng, and N. Zheng, "Fine-grained image classification using modified DCNNs trained by cascaded softmax and generalized large-margin losses," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 683–694, Mar. 2019.

[4] 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.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. IEEE Int. Conf. Mach. Learn. (ICML)*, Jul. 2015, pp. 2048–2057.

[6] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, 2015, Art. no. e0130140.

[7] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4307–4314.

[8] A. Mahendran and A. Vedaldi, "Salient deconvolutinoal networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 120–135.

[9] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[11] J. Yosinski, J. Clune, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Proc. ICML Workshop Deep Learn.*, Jun. 2015, pp. 1–12.

[12] M. Robnik-Sikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 589–600, May 2008.

[13] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.

[14] V. Petisiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Jun. 2018, pp. 1–17.

[15] D. Seo, K. Oh, and I.-S. Oh, "Regional multi-scale approach for visually pleasing explanations of deep neural networks," *IEEE Access*, vol. 8, pp. 8572–8582, 2020.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. SIGKDD*, Aug. 2016, pp. 1135–1144.

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

[18] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2017, pp. 1–12.

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

[20] J. T. Springenberg, A. Dosoviskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. ICLR Workshop DeepLearn.*, Apr. 2015, pp. 1–14.

[21] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

[22] S. Tian and C. Ying, "Visualizing deep neural networks with interaction of super-pixels," in *Proc. Conf. Inf. Knowl. Mgmt.*, Dec. 2017, pp. 2327–2330.

[23] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1988–1996.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[25] C. Qiu, S. Zhang, C. Wang, Z. Yu, H. Zheng, and B. Zheng, "Improving transfer learning and squeeze-and-excitation networks for small-scale fine-grained fish image classification," *IEEE Access*, vol. 6, pp. 78503–78512, 2018.

[26] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1487–1500, Mar. 2018.

[27] X. Li, L. Yu, D. Chang, Z. Ma, and J. Cao, "Dual cross-entropy loss for small-sample fine-grained vehicle classification," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4204–4212, May 2019.

[28] Z. Lin, S. Mu, F. Huang, K. A. Mateen, M. Wang, W. Gao, and J. Jia, "A unified matrix-based convolutional neural network for fine-grained image classification of wheat leaf diseases," *IEEE Access*, vol. 7, pp. 11570–11590, 2019.

[29] R. Sipes and D. Li, "Using convolutional neural networks for automated fine grained image classification of acute lymphoblastic leukemia," in *Proc. 3rd Int. Conf. Comput. Intell. Appl. (ICCIA)*, Jul. 2018, pp. 157–161.

[30] H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3713–3717.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep., 2011.

[32] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.

[33] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*. [Online]. Available: http://arxiv.org/abs/1306.5151

[34] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, Jun. 2014, pp. 675–678.

[35] K. Oh, W. Kim, G. Shen, Y. Piao, N.-I. Kang, I.-S. Oh, and Y. C. Chung, "Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization," *Schizophrenia Res.*, vol. 212, pp. 186–195, Oct. 2019.

[36] L. Zhou, Y. Wang, Y. Li, P.-T. Yap, and D. Shen, "Hierarchical anatomical brain networks for MCI prediction: Revisiting volumetric measures," *PLoS ONE*, vol. 6, no. 7, Jul. 2011, Art. no. e21935.

[37] K. H. Oh, Y.-C. Chung, K. W. Kim, W.-S. Kim, and I.-S. Oh, "Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning," *Sci. Rep.* vol. 9, Dec. 2019, Art. no. 18150.

[38] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, Jan. 2002.

KANGHAN OH received the B.S. degree in computer science from Honam University, South Korea, in 2010, and the Ph.D. degree in electronic and computer engineering from Chonnam National University, South Korea, in 2017. He is currently a Postdoctoral Researcher with the Division of Computer Science and Engineering, Jeonbuk National University, South Korea. His research interests include computer vision, pattern recognition, and machine learning.

SUNGCHAN KIM received the B.S. degree in material science and engineering, the M.S. degree in computer engineering, and the Ph.D. degree in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1998, 2000, and 2005, respectively. He is currently a Professor with the Division of Computer Science and Engineering, Jeonbuk National University, South Korea. His research interests include various topics in machine learning and computer vision.

IL-SEOK OH received the B.S. degree in computer engineering from Seoul National University, South Korea, in 1984, and the Ph.D. degree in computer science from KAIST, South Korea, in 1992. He is currently a Professor with the Division of Computer Science and Engineering, Jeonbuk National University, Jeonju, South Korea. He was a Visiting Scientist with CENPARMI, Concordia University, Canada, and UCI, USA. He is the author of the books *Pattern Recognition*, *Computer Vision*, and *Machine Learning* (Korean Language). His research interests include computer vision, pattern recognition, and machine learning.