GSM-HM: Generation of Saliency Maps for Black-Box Object Detection Model Based on Hierarchical Masking

YICHENG YAN1, XIANFENG LI1, (Member, IEEE), YING ZHAN1, LIANPENG SUN2, AND JINJUN ZHU2

1School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China
2School of Environmental Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

Corresponding author: Xianfeng Li (xifli@must.edu.mo)

This work was supported in part by the Science and Technology Development Fund of Macau under Grant 0079/2019/AMJ, and in part by the National Key Research and Development Program of China under Grant 2019YFE0111400.

ABSTRACT Interpretablility of DNN-based object detection has been a rising concern for the research community. The first step towards this goal is a saliency map that visualizes the importance (saliency) of pixels in an image for the object detected by a specific model. Black-box based methods generate a saliency map without the need to look into the internals of a model, thus applicable to all models without the need of adaptation. In addition, they provide more reliable evaluation on the saliency of pixels than white-box methods by means of the absence of these pixels from the image. However, with current black-box methods, the absence of pixels is produced by random image masks. Despite the need of a great number of random masks for sufficient coverage, the quality of the pixel saliency is not assured to be satisfactory. In this work, we propose a more effective black-box framework with hierarchical masking. In this framework, called GSM-HM, pixel saliency is evaluated at multiple levels, with each lower level performing a refinement on the saliency information of the upper level. This hierarchical framework significantly reduces the masking efforts on less valuable pixels, thus it can produce saliency maps with higher qualities. In our experiments, the quality of a generated saliency map is evaluated with four different metrics: deletion, insertion, convergence and RAM (the ratio of average to maximum). Compared with D-RISE, a recent black-block method, GSM-HM generates more accurate saliency maps evaluated by these metrics.

INDEX TERMS Saliency map, black-box model, object detection, explainable artificial intelligence.

I. INTRODUCTION

Object detection is one of the important and challenging tasks in computer vision. Its main task is to detect the category, size, location and other information of an object in an input image or video [1]. With the development of deep learning technology, object detection algorithms gradually evolve from traditional algorithms (such as HOG [2], etc.) to algorithms based on deep learning [3] (such as the famous one-stage detection model YOLO [4], two-stage detection model Faster R-CNN [5], etc.). Algorithms based on deep learning can not only extract object features automatically, but also improve the accuracy of detection. At present, DNN-based object detection algorithms have been applied to different tasks because of its excellent performance.

However, due to the poor interpretability of deep neural networks, it is difficult to modify and optimize the DNN-based object detection model. In order to better understand deep neural networks, a lot of work has been done to observe deep neural networks through visualization techniques [6], such as the visualization method proposed in [7] and the activation maximization proposed in [8]. By these visualization technologies, the features extracted and learned by different layers in deep neural networks can be obtained, which can help us to better understand neural networks.
Among the methods for explaining DNN, the saliency map has become a popular tool. A saliency map is essentially a heat map, where different pixels in an image exhibit different influences on the detection of a specific object. At present, there are two main approaches to saliency map generation. One is to generate saliency maps based on the gradients of deep neural networks, such as Grad-CAM [9]. The other is to regard DNN-based object detector as a black box, and to obtain saliency areas related to an object by applying perturbation information to the input image, such as D-RISE [10].

The black-box approach has two advantages over the gradient-based approach. Firstly, it does not need to explore the internal structure of a specific model, and is more applicable to different models. More importantly, since the saliency scores of pixels in the black-box methods are based on perturbations (e.g., masking an area) to the input image, the saliency scores of the pixels are evaluated from their absence from the image. In contrast, gradient-based methods can only quantify the contribution of the pixels to object detection at their presence. It is possible that some high-gradient pixels are actually not so important for the object if the gradients of some other pixels get increased with the absence of the original high-gradient pixels. In other words, the absence of pixels instead of their presence provides a more reliable means for evaluating their saliency for object detection. Because of the above reasons, our work pursues the black-box approach.

D-RISE [10] is the first black-box saliency map method for detectors. It first divides the image into grids of cells in fixed size, then a set of masks are generated by randomly masking some cells in each mask. Next, each mask is applied to the input image to generate a masked image, which will be fed into the model to obtain its score of object detection. From the difference of scores between the original image and the masked image, it obtains a saliency score for each mask. This saliency score will also be assigned to each masked cell in the corresponding mask. The final saliency of each cell is obtained from the sum of its scores across the set of masks, and the saliency map of the image is generated accordingly.

However, since D-RISE adopts the random masking strategy, the low-saliency areas can also obtain a high score when some high-saliency areas and low-saliency areas appear in a mask simultaneously, which will introduce additional "noise" into the saliency map. Moreover, the random masking strategy may generate non-smooth masking areas, which may introduce additional adversarial artifacts as analyzed in [11]. Although these adversarial artifacts do not affect human detection, they will have a devastating effect on the detector. These adversarial artifacts result in the poor interpretation of the generated saliency map.

In this paper, we propose a black-box saliency map generation method with hierarchical masking, which updates the saliency scores of pixels hierarchically and refines the saliency map smoothly. Our contributions are as follows:

- We design a hierarchical masking framework consisting of a coarse-grained phase that identify the approximate saliency areas for an object, and a fine-grained phase for refining the saliency scores of these areas.
- We propose an adaptive mask generation mechanism using the $l$-nearest neighbors. It can adapt to objects of different sizes automatically, depending on whether the object is large or small.
- We propose a new evaluation metric to better evaluate the quality of saliency maps.
- Compared with D-RISE, the saliency maps generated by our hierarchical framework have less "noise" and better represent the saliency areas of objects.

II. RELATED WORK

Saliency maps can be a useful tool for understanding, evaluating and optimizing object detection models. There exist two basic approaches for generating saliency maps.

A. GRADIENT-BASED APPROACH

In gradient-based methods, the saliency map is calculated according to the gradients in the model. Simonyan et al. proposed a gradient-based method named Gradient [12]. This method obtains the derivatives of each pixel in the input image by backward propagation, and then rearranges the derivative vector to obtain the saliency map of the input image. For multi-channel images, Gradient will take the maximum derivative for each pixel across all channels. CAM [13] replaces the full connection layer in the classifier with the global average pooling layer, and obtains the weight of each feature map after global average pooling layer according to softmax results. Then class activation map is generated by a weighted sum of these feature maps. However, CAM has to modify the architecture of model, which imposes additional overhead on model training. So a general method, Grad-CAM [9], was proposed. Grad-CAM calculates the average gradient of the last convolutional layer as the weight, and sums the feature maps of the last layer according to the weight to obtain the saliency map. There are other methods based on CAM, such as Grad-CAM++ [14]. Based on Grad-CAM, LayerCAM [15] was proposed to generate saliency map of every layer of convolutional neural networks. LayerCAM can obtain more fine-grained salient information of an object by class activation maps of shallow layers.

As explained earlier, this approach has the problem that class activation map is not equivalent to saliency map, and the absence of the area with strong activation might have no appreciable effect on object detection.

B. BLACK-BOX APPROACH

Black-box methods are not based on gradient information, thus no need to examine the internals of the deep models. Instead, the impact of an area on the object detection is evaluated by masking it off the input image. This makes the black-box methods applicable to all deep models without any adaptation. The saliency map generated by it would also be more objective as the impact for object detection is evaluated from the absence of an area.
LIME [16] is one of the representative methods based on black-box. It takes a random masking sample around the input object to get a new set of samples. Then, according to the distance between new samples and the original sample, a simple linear model is fitted to explain the model. RISE [17] is a method to explain classification models by random masking. It randomly generates some masks to modify the original input image and calculates the score of the masked image. Then the score as the saliency will be sent to the masked cells to generate an individual saliency map. The final saliency map is obtained by superimposing multiple individual saliency maps. D-RISE [10] explains detectors on the basis of RISE by modifying the scoring method of masked image. Mokuwe et al. [18] proposed a saliency map generation method based on Bayesian optimization, which is used to optimize the generation of masks for more accurate saliency maps. A morphological fragmental perturbation pyramid (MFPP) [19] method was proposed to solve the interpretability problem. This method can divide the original image into morphological fragments of different scales and randomly mask some fragments. This method can fully consider the input semantic information and improve the accuracy of saliency map.

At present, most black-box methods adopt an uncontrolled random masking mechanism, which often bring together irrelevant pixels with negative effects. Although a large number of masks are evaluated to reduce the negative effect, it still introduces saliency pollution on the map. To overcome this problem, a mechanism that can efficiently cover salient pixels with controlled randomness is needed.

III. METHODOLOGY
A. TERMINOLOGY
To facilitate detailed description in the rest of this paper, we first introduce several common terms and their notations used in this paper.

- Cell ($c$): In our hierarchical masking framework, an image is partitioned into equal-sized cells at each level. The number of pixels within a cell can be calculated from the corresponding level of the cell. The cell is also the smallest unit to make up a mask at a level. For example, an $8 \times 8$ mask covers 64 cells.
- Mask ($M$): $M$ contains the state of each cell in the image. If a cell $c$ is masked, it is denoted as $M(c) = 1$, otherwise $M(c) = 0$.
- Masked Image ($I'$): $I'$ is obtained by applying a mask $M$ to the original image $I$. Some pixels are masked in $I'$ which are set to 0.
- Saliency Score ($SS$) of a mask: $SS$ reflects the impact of a mask on object detection. It is calculated from the difference on object detection results on the original image $I$ and the masked image $I'$.
- Saliency Value ($SV$) of a cell: $SV$ is calculated by all mask saliency scores.

B. BASELINE BLACK-BOX FRAMEWORK
For comparison, we describe a baseline black-box framework for saliency map generation, which is essentially the framework used in D-RISE [10]. Suppose we have a DNN-based object detector $D$ that can output bounding boxes, labels and confidence scores of detected objects for a given image $I$, our goal is to generate a saliency map for each of the detected objects in this image.

First, the input image $I$ is perturbated by a mask $M$ that covers some pixels of the image, and the masked image $I'$ will be tested by the object detector $D$. For a specific object $O$ in image $I$, we obtain the original confidence score $CS_I$ and the masked confidence score $CS_{I'}$ with the masked image $I'$ respectively, as shown in Fig. 1. In addition, we also obtain the IoU of the bounding boxes in $I$ and $I'$ respectively, which is denoted as $IoU_{I,I'}$ in Fig. 1. The Saliency Score of mask $M$, $SS_M$, can be calculated with the following formula:

$$SS_M = 1 - \frac{IoU_{I,I'} \times CS_{I'}}{CS_I} \quad (1)$$

Essentially, the larger the $SS_M$, the more likely that at least part of the masked pixels are salient for object $O$. Thus we can generate a saliency map specific to mask $M$, where the masked pixels will be assigned the saliency score $SS_M$.

However, a single saliency map obtained from a specific mask has two drawbacks. On one hand, it may miss some other salient pixels that are not covered by the mask. On the other hand, some masked pixels might actually not be salient for this object. They are assigned a high salient score simply because they happen to be masked together with other salient pixels. Therefore, in the baseline framework, a large set of diversified masks are randomly generated and applied to
the object detector. The saliency maps generated from these masks are then fused to obtain the final saliency map for the corresponding object in the image.

**C. HIERARCHICAL FRAMEWORK**

With the baseline framework, each mask is randomly generated without exploiting any saliency information obtained from other masks. Such uncontrolled randomness often brings together irrelevant pixels in a mask. It not only has a consequence on less effective masks, but also causes saliency pollution in areas where non-salient pixels receive high saliency scores because they happen to be masked together with salient pixels for certain times. But on the other hand, without randomness, the search for salient pixels might be optimistically restricted to certain areas, leading to a partial saliency map that fail to cover many salient pixels.

To guide the search for salient pixels effectively and safely, we propose a framework for the Generation of Saliency Maps based on Hierarchical Masking (GSM-HM). The basic idea of this framework is to leverage the coarse-grained saliency information at the upper levels for the generation of more refined saliency maps at the lower levels. Fig. 2 shows the GSM-HM framework using the truck object as an example. The number of cells in a mask at level \( l + 1 \) is four times that at level \( l \). At each level, a set of masks are generated based on the cell saliency values inherited from the previous level.

Next, the masked images are fed into the object detector to obtain the saliency scores for the corresponding masks. These mask saliency scores are then used to update the cell saliency values. When this process terminates at the lowest level, the final saliency map is generated. GSM-HM is divided into a coarse-grained phase, followed by a fine-grained phase, each having several levels. At coarse-grained levels, we obtain a relatively accurate prior saliency knowledge of cells. At fine-grained levels, the saliency area of an object is continuously refined.

**D. MASK GENERATION**

The first step in GSM-HM is mask generation at a level \( l \), and its algorithm is described in this part.

1) \( l \)-NEAREST NEIGHBORS MASK

Based on the fact that salient pixels for an object are naturally formed into a set of continuous areas of different sizes, we propose a \( l \)-nearest neighbors mask generation mechanism. For each mask at level \( l \), a cell is first selected as the center of the masking area with a probability weighted on its saliency value \( SV \) (which means cells with higher saliency values will have higher chances). Next, the \( l \)-nearest neighbors of the center cell are selected to form the whole masking area. Fig. 4(b) shows that the blue cell is the center cell of the masking area, and the yellow cells are the neighbors.
around areas with higher saliency. This means that we have
derived from previous levels, will guide the generation of masks
for center cell selection, which is based on saliency information.
We design different mask generation termination conditions
for the coarse-grained and fine-grained levels. Generally, the mask
generation termination condition is to set a fixed number of masks.
For example, the number of masks set by D-RISE [10] is 5000. However,
since the properties of objects such as size and shape are different,
a fixed number of masks may not necessarily achieve ideal performance
for all objects. In order to improve the adaptability of masks to
the properties of objects, we design the mask generation termination condition
based on the sum of cell saliency values for the cell candidate set $\mathcal{S}_C$, as shown in (2).

$$SV_{\text{SUM}}_C = \sum_{c \in C} SV_c$$

At the coarse-grained levels, we hope to obtain relatively accurate prior knowledge about the cell saliency, so the mask
generation termination condition of coarse-grained levels is
that the center cell candidate set $C$ becomes $\emptyset$.
At the fine-grained levels, due to a large number of cells, a large number of masks need to be generated if all cells in
$C$ are selected as the center cell. To reduce the number of

Another difference to the baseline framework is that the
masked areas at different levels are different instead of a constant size. In particular, the number masked cells at level $l$ is $(2l+1)^2$ (with the top level $l = 0$). Since the size of a cell at
level $l=1$ is $1/4$ of a cell at level $l$, the masking area at level $l$ is $(2l+1)^2/4^l$ of the size at level 0. Clearly, the masking area is shrinking with the increasing of levels, although the number
of masked cells are increasing. Their contrast is plotted in
Fig. 3. As a concrete example, Fig. 4 compares the masked
cells and areas between level 1 and level 2. A mask at $l = 1$ on the left side covers an area of 9 units which is equivalent to
an area of $6 \times 6 = 36$ units at $l = 2$, whereas a mask in the same place at $l = 2$ on the right side covers an area of
$5 \times 5 = 25$ units, although it has more masked cells (with a ratio of 25/9).

The increasing on the number of masked cells and the
shrinking of masked area, together with the weighted selection
mechanism for center cell selection, will facilitate generating high quality saliency maps for both large objects and
small objects due to the following reasons. First, the weighted
selection mechanism, which is based on saliency information
from previous levels, will guide the generation of masks around areas with higher saliency. This means that we have
more chances to refine the shapes of salient areas at lower levels. Second, at higher levels, the masked areas with large sizes
can very quickly identify potential saliency areas, although they may have a significant portion of non-salient pixels for small objects. Last, at lower levels, the masked areas consist of smaller cells, which will help refine the saliency areas identified at upper levels. In particular, the refining of the saliency areas more often results in confirmation of saliency in case of large objects, and more often results in exclusion of saliency in case of small objects.

2) GENERATION PROCESS
The process of the $l$-nearest neighbors mask generation algo-

**Algorithm 1** Algorithm for Mask Generation at Level $l$

- **Input:** $C^{(l)}$: the set of cells at level $l$
- **while** $MG_{\text{Finish}}(l) = \text{False}$
  1. $C = \{c | c \in C^{(l)} \land SV_c > 0\}$
  2. $B \leftarrow$ Weighted randomly sample 10 cells from $C$
  3. $M_i = \{M = \text{Nearest}_\text{Neighbors}(c), \forall c \in B\}$
  4. $M_c \leftarrow \text{Best}_\text{Mask}(M_i)$,
     where $c$ is the selected center for mask $M_c$
  5. Remove $c$ from $C$
  6. **end while**

At each level, cells with non-zero saliency values are put into $C$ as candidates of center cells. Note that level 0 is a special level because this level doesn’t have the previous level. So all cells at level 0 are set to the same cell saliency value and put into the center cell candidate set $C$. According to cell saliency values, the center cell is weighted randomly sampled from the $C$. When a cell is chosen as the center cell, it will be removed from $C$. Then, the $l$-nearest neighbors mask is generated based on the center cell and its neighbors.

**FIGURE 3.** The number of masked cells and area at each level. The columns show the number of masked cells at each level. The red line shows the masked area ratio at each level compared to level 0.

**FIGURE 4.** The change of mask from level 1 to level 2. The colored cells are masked.
masks at each of these levels, we set a termination condition according to the following formula:

\[
\frac{SV_{\text{SUM}}_{c'}}{SV_{\text{SUM}}_c} \leq \delta
\]

(3)

where \(C'\) contains the rest cells that have not been selected as mask centers yet. The rationale is that if the sum of saliency scores for the rest of the candidate cells in \(C'\) is below a certain level, the need of trying more masks becomes low.

E. MASK SALIENCY SCORE

As shown in Fig. 2, we can evaluate the Saliency Score (SS) for each mask by comparing the difference on object detection results between the original image and the corresponding masked image. The calculation of SS has already been given in (1) when describing the baseline model in Section III-B.

Note that \(SS_M \in [0, 1]\). When a mask \(M\) has little influence on object detection, its \(SS_M\) is close to 0. In contrast, when the mask has a significant influence on object detection, \(SS_M\) is close to 1.

Since an input image may contain multiple objects with the same label, we adopt the following strategy to reduce their interference: for all results output by the model with the same object label, the one with the largest IoU is taken for the calculation of saliency score for the corresponding object in the masked image.

F. CELL SALIENCY VALUE

As shown in Fig. 2, the Saliency Value (SV) of a cell can be calculated from the Saliency Scores (SS) of the masks at each level. In essence, \(SV_c\) of a cell \(c\) is calculated as the average of Mask Saliency Scores where \(c\) is masked. Let the set of masks at level \(l\) be \(M_{\text{or}}(l)\). For each mask \(M \in M_{\text{or}}(l)\), \(M(c) = 1\) means cell \(c\) is masked in \(M\), otherwise \(M(c) = 0\). Therefore, we use (4) to calculate the Saliency Value for each cell \(c\) at level \(l\).

\[
SV_c = \frac{\sum_{M \in M_{\text{or}}(l)} M(c) \times SS_M}{\sum_{M \in M_{\text{or}}(l)} M(c)}
\]

(4)

In addition to Saliency Value \(SV_c\) evaluated at current level \(l\), we also consider the result \(SV_c^{(l-1)}\) inherited from the parent cell at the previous level. The two parts are combined to get \(SV_c^{(l)}\), the final Saliency Value for cell \(c\) at level \(l\), as given by the first case in (5).

\[
SV_c^{(l)} = \begin{cases} 
\alpha \times SV_c + (1 - \alpha) \times SV_c^{(l-1)}, & \exists M, M(c) = 1 \\
SV_c^{(l-1)}, & \text{otherwise}
\end{cases}
\]

(5)

The parameter \(\alpha\) is a weight for deciding which part has more impact on the final result at level \(l\). When \(\alpha\) approaches 1, the saliency value inherited from the previous level is almost neglected, and when \(\alpha\) becomes smaller, more impact from the previous level is accounted. But if a cell \(c\) is not covered by any mask at level \(l\), \(SV_c\) will become 0 according to (4). In this case, we let \(SV_c^{(l)} = SV_c^{(l-1)}\), as given by the second case in (5). This case means that cell \(c\) completely inherits the Saliency Value of its parent cell at level \(l - 1\), because it has no chance to get its Saliency Value updated at level \(l\) by being tested with masks.

G. SALIENCY MAP GENERATION

The Saliency Map at a level can be obtained straightforwardly from the Saliency Values of cells of that level. But according to our ultimate goal, the Saliency Map at the last level is of our interest, as it gives the most detailed saliency values of the image.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENT SETTINGS

We use YOLOv5 implemented by PyTorch as the detector and randomly sample some images from the MS-COCO dataset as our test dataset. In our hierarchical framework GSM-HM, the image at the top level (level 0) is partitioned into 8 \(\times\) 8 equal-sized cells. There are totally five levels in our hierarchy, and the four levels below the top level are partitioned into cells of 16 \(\times\) 16, 32 \(\times\) 32, 64 \(\times\) 64, 128 \(\times\) 128, respectively. Each lower level contains more cells of smaller sizes, thus allowing more fine-grained saliency evaluation. In particular, we treat 8 \(\times\) 8, 16 \(\times\) 16, and 32 \(\times\) 32 as coarse-grained levels, and the rest two (64 \(\times\) 64 and 128 \(\times\) 128) as fine-grained levels.

The termination condition of the mask generation process are different for coarse-grained and fine-grained levels, as described in Section III-D. At each fine-grained level, we set the mask generation termination threshold \(\delta = 0.5\). We set the weight for cell saliency calculation \(\alpha = 0.5\). D-RISE [10] is used as the comparison method.

B. EVALUATION METHODS

In order to measure the quality of saliency maps generated by different techniques, this paper adopts four different evaluation methods, namely Detection [10], Insertion [10], Convergence [20], and RAM (the ratio of average to maximum). These four methods can be used to evaluate the saliency map from different aspects.

Detection and Insertion, used in D-RISE, can evaluate the correlation between the saliency areas in the saliency map and the object. In other words, whether the saliency area has a significant impact on the detection of objects. Detection deletes pixels from the original image one by one in descending order of saliency and measures the difference between the detection result of the original image \(I\) and the result of the image \(I'\) with deleted pixels. This difference of the detection results, denoted as \(DDR_{I,I'}\), can be evaluated by the following formula:

\[
DDR_{I,I'} = \frac{\text{IoU}_{I,I'} \times CS_{I'}}{CS_I}
\]

(6)

The right part of the formula was also used for calculating the Saliency Score of a mask in (1). The drop speed of \(DDR_{I,I'}\) is faster if the correlation between the saliency area and the object is stronger. Different from deletion, insertion fills the pixels into the original image one by one in...
Y. Yan et al.: GSM-HM: Generation of Saliency Maps for Black-Box Object Detection Model

FIGURE 5. Saliency maps at different levels of GSM-HM (our). The objects in these saliency maps are: parking meter, train and baseball bat.

descending order of saliency and measures $DDR_{I,I'}$. The growth speed of $DDR_{I,I'}$ is faster if the correlation between the saliency area and the object is stronger.

Convergence is used to evaluate the degree of convergence of saliency maps. The black-box based approach usually adopts a random masking method, so it is very important that it can output relatively stable saliency maps. For each work, three experiments are performed to get three saliency maps from the same input image with the same experimental settings. Then the convergence is measured by calculating the Euclidean distance between each pair of the three saliency maps, as shown in formula (7).

$$Convergence = \frac{\|SM_{1,2}\| + \|SM_{1,3}\| + \|SM_{2,3}\|}{3} \tag{7}$$

where $\|SM_{1,2}\|$ means the Euclidean distance between saliency maps $SM_1$ and $SM_2$. The smaller the Euclidean distances, the better the convergence of the work.

RAM (the ratio of average to maximum) proposed in this paper is motivated by observations on the saliency map results of different algorithms. As shown in the first row of Fig. 6, both our GSM-HM framework and D-RISE can highlight the saliency area for the tennis racket object, but some irrelevant areas are also assigned non-trivial saliency by D-RISE. We introduce (8) to evaluate the influence of this kind of “noise” on the saliency map, where $\text{Avg}(SMAP)$ and $\text{Max}(SMAP)$ are the average and maximum saliency values of a saliency map $SMAP$ respectively. The more “noise” in the saliency map, the higher the value of $\text{Avg}(SMAP)$, and the higher the RAM metric. Therefore, a good framework for saliency map is expected to have a low RAM metric.

$$\text{RAM} = \frac{\text{Avg}(SMAP)}{\text{Max}(SMAP)} \tag{8}$$

C. EXPERIMENTAL RESULTS

In GSM-HM, a saliency map is generated at each level of granularity, as shown in Fig. 5. At the initial level, we can obtain some rough saliency areas of an object and exclude some irrelevant areas. Compared to non-hierarchical random methods like D-RISE, this will help reduce efforts spent on a large portion of the image for saliency evaluation at lower levels, which have increasingly larger number of cells to test.

Fig. 6 gives the saliency maps generated by GSM-HM and D-RISE for several objects of different sizes and shapes,
TABLE 1. The mean value of the four evaluation metrics.

| Method      | Mean Deletion | Mean Insertion | Mean Convergence | Mean RAM  |
|-------------|---------------|----------------|------------------|----------|
| GSM-HM      | 0.042         | 0.851          | 5.448            | 0.096    |
| D-RISE      | 0.091         | 0.636          | 16.793           | 0.310    |

respectively. We have three major observations. The first observation is that both methods can find a similar saliency area of the object. But unlike D-RISE, there is very little noise in the saliency map generated by GSM-HM. The reason is apparent: with D-RISE, many irrelevant pixels are tested together with real salient pixels due to the random generation of masks, thus these irrelevant pixels will get an inappropriate high value of saliency. GSM-HM can greatly reduce testing a large portion of irrelevant areas with the saliency knowledge of their parent cells from the upper level.

The second observation is that although the saliency areas are similar, their degrees of significance are different, and the salient areas seem to be more concentrated in GSM-HM. For example, the shaft of the tennis racket and the legs of the bird are more salient than other parts of their corresponding objects in GSM-HM. This kind of sharper contrast will help identify the more important parts for object detection.

The third observation is that the saliency maps generated by D-RISE appear to be smoother. But the smoothness of D-RISE is an artificial effect applied to the original saliency map with a simple math operation. The smoothness of our work is purely the result of the hierarchical refining of saliency.

Table 1 gives a comparison between GSM-HM and D-RISE on the mean value for each of the four evaluation metrics with the test dataset. It can be seen that GSM-HM is consistently better than D-RISE. More in-depth analysis of the four metrics will be described in the rest part.

D. ANALYSIS

1) DELETION AND INSERTION

As discussed earlier, deletion and insertion can be used to evaluate the accuracy of saliency areas. In particular, when we get the curve of the deletion and insertion, the area under the curve (AUC) can be calculated for evaluation. As shown in Table 1, the mean deletion AUC value of GSM-HM is lower than D-RISE, and its mean insertion AUC value is higher than D-RISE, both are in favor of GSM-HM.

Fig. 7 gives a concrete example on the deletion and insertion results. With GSM-HM, the deletion curve drops sharply and the insertion curve grows faster. The main reason is that unlike D-RISE, our GSM-HM framework less frequently assigns irrelevant areas with inappropriately high saliency.

2) CONVERGENCE

Since the magnitude of saliency maps generated by different methods are different, we first normalize the saliency maps.

Then the convergence metric is calculated with formula (7). From the mean convergence value shown in Table 1, we can see that the convergence of GSM-HM is considerably better.

Table 2 gives some examples of convergence results. Each pair of D-RISE results have a larger Euclidean distance, which means the saliency areas of different results are significantly different. The poor convergence performance of D-RISE is caused by the high saliencies of irrelevant areas. The example in Table 3 further illustrates the reason for the poor convergence of D-RISE. It can be seen that areas far from the motorcycle object have different degrees and distributions of saliencies, and this kind of difference causes the poor performance on the convergence metric of D-RISE. In contrast, in our GSM-HM framework, salient areas are refined with the guidance of prior knowledge obtained at coarse granularities, and the salient areas are closely restricted to the object itself, thus it achieves a much better convergence by excluding irrelevant areas from object saliency.

3) RAM

When calculating the RAM metric with formula (8), a max-min normalization of the saliency map is needed, such that the results of different methods are in the same magnitude.

It can be seen from Table 1, the RAM value of D-RISE is higher than GSM-HM. As described earlier, the RAM metric is introduced to study the “noise” on saliency. As expected, the “noise” in the saliency map generated by D-RISE leads to high RAM value. Table 4 gives a concrete example for the truck object in the image. With D-RISE, the area outside
V. CONCLUSION

In order to generate a high-quality saliency map for black-box detectors, this paper proposes a saliency map generation framework based on hierarchical image masking. The framework, called GSM-HM, addresses the issues caused by random mask generation methods. The saliency map generated by GSM-HM is sharper than D-RISE, the state-of-the-art black-box method. GSM-HM also generates saliency maps with less “noises” than D-RISE. We also perform quantitative comparisons with D-RISE using four metrics. Experimental results demonstrate that our method is able to generate more accurate object-related saliency maps.

In future work, how to generate accurate saliency maps with significant less masks is a research topic that deserve investigating. Current masking method also needs to be improved, as it may introduce new artifacts when covering an area with black pixels. We also have a strong interest in using the saliency maps as a tool for investigating some tricky issues raised in object detection.

REFERENCES

[1] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” May 2019, arXiv:1905.0505.
[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 886–893.
[3] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, “A survey of deep learning-based object detection,” IEEE Access, vol. 7, pp. 128837–128868, 2019.
[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[6] Z. Qin, F. Yu, C. Liu, and X. Chen, “How convolutional neural network see the world—A survey of convolutional neural network visualization methods,” 2018, arXiv:1804.11191.
[7] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 818–833.

[8] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” Univ. Montreal, Montreal, QC, Canada, Tech. Rep. 1341, 2009.

[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,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.

[10] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, “Black-box explanation of object detectors via saliency maps,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 11443–11452.

[11] P. Dubowsky and Y. Gal, “Real-time image saliency for black box classifiers,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2017, pp. 6970–6979.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” 2013, arXiv:1312.6034.

[13] 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.

[14] 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.

[15] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, “LayerCAM: Exploring hierarchical class activation maps for localization,” IEEE Trans. Image Process., vol. 30, pp. 5875–5888, 2021.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in Proc. SIGKDD, 2016, pp. 1135–1144.

[17] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized input sampling for explanation of black-box models,” in Proc. BMVC, 2018, pp. 1–17.

[18] M. Mokuwe, M. Burke, and A. S. Bosman, “Black-box saliency map generation using Bayesian optimisation,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2020, pp. 1–8.

[19] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, “MFPFP: Morphological fragmental perturbation pyramid for black-box model explanations,” in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 1376–1383.

[20] L. Brunke, P. Agrawal, and N. George, “Evaluating input perturbation methods for interpreting CNNs and saliency map comparison,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 120–134.

YICHENG YAN received the M.S. degree in software engineering from Northeastern University, Shenyang, Liaoning, China, in 2021. He is currently pursuing the Ph.D. degree in advanced networking with the Macau University of Science and Technology, Macau, China.

His current research interests include edge intelligence, edge computing, and computer vision.

YING ZHAN received the M.S. degree in communication and information system from the Communication University of China, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree in artificial intelligence with the Macau University of Science and Technology, Macau, China.

Her current research interests include object detection, computer vision, and emotion recognition.

LIANPENG SUN received the B.Sc. degree in water supply and sewage engineering from the Harbin University of Civil Engineering and Architecture, and the Ph.D. degree in municipal engineering from the Harbin Institute of Technology, in 2001.

He has been the Dean of the School of Environmental Science and Engineering, Sun Yat-sen University, since 2020, and the Director of the Guangdong Water Pollution Control Engineering Technology Research Center, since 2016. His current research interests include drainage pipe network informationization, pollutant traceability technology sludge treatment, and sludge reduction technology.

JINJUN ZHU received the B.Sc. degree in environmental engineering from Sun Yat-sen University, in 2021, where she is currently pursuing the master’s degree in environmental engineering.

Her current research interest includes drainage pipe network informationization.

XIANFENG LI (Member, IEEE) received the B.Sc. degree from the Beijing Institute of Technology, China, in 1995, and the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2005.

He is currently an Associate Professor with the Macau University of Science and Technology. His current research interests include edge intelligence, smart Internet of Things, high-performance networking, and hardware/software codesign.