APPENDIX

In this appendix, we provide the proof for Theorem 1 in Section A, and provide more experiment results to demonstrate the outstanding interpretation performance of RI in Sections B-G.

A. The Proof of Theorem 1

We first rewrite Equation (3) as

$$\min_{P(x)\subseteq\mathcal{Q}} |R| - |P(x)\cap R| \tag{9a}$$

$$\text{s.t. } |D(x)| - |P(x)\cap D(x)| \geq |D(x)| - \delta, \tag{9b}$$

where $|R|$ and $|D(x)|$ are the constant cardinalities of R and $D(x)$, respectively.

Then, we prove the problem of Equation (9) is an SCSC problem [28] by showing

$$|R| - |P(x)\cap R| \tag{10}$$

and

$$|D(x)| - |P(x)\cap D(x)| \tag{11}$$

are submodular functions with respect to $P(x)\subseteq\mathcal{Q}$.

Denote by $P = P(x)$, and by $f(P) = |R| - |P\cap R|$. We prove $f(P)$ is a submodular function with respect to $P\subseteq\mathcal{Q}$ by showing that, for any two sets of linear boundaries, denoted by $P\subseteq\mathcal{Q}$ and $T\subseteq\mathcal{Q}$, if $P\subseteq T$, then

$$f(P\cup\{\mathbf{h}\}) - f(P) \geq f(T\cup\{\mathbf{h}\}) - f(T) \tag{12}$$

holds for any linear boundary $\mathbf{h}\in\mathcal{Q}\setminus T$.

Recall that the linear boundaries in P defines a convex polytope, and $|P\cap R|$ is the number of images in R that are covered by P. Thus, $f(P) = |R| - |P\cap R|$ is the number of images in R that are not covered by P.

If we add a new linear boundary $\mathbf{h}\in\mathcal{Q}\setminus T$ to P, some images covered by P may not be covered by the new convex polytope defined by $P\cup\{\mathbf{h}\}$. We say these images are removed by $\mathbf{h}$ from P.

The left side of Equation (12) is exactly the number of the images removed by $\mathbf{h}$ from P. Similarly, the right side of Equation (12) is the number of the images removed by $\mathbf{h}$ from T.

Since $P\subseteq T$, the set of images covered by P contains the set of images covered by T. Therefore, the number of images removed by $\mathbf{h}$ from P will be no smaller than the number of images removed by $\mathbf{h}$ from T. As a result, Equation (12) holds, which means $|R| - |P(x)\cap R|$ is a submodular function with respect to $P(x)$.

We can prove $|D(x)| - |P(x)\cap D(x)|$ is a submodular function with respect to $P(x)$ in the same way. This concludes the theorem.

B. Case Study

We present a case study on the FOOD dataset in Figure 3. The experiment setting is the same as the case study discussed in Section 6.3.

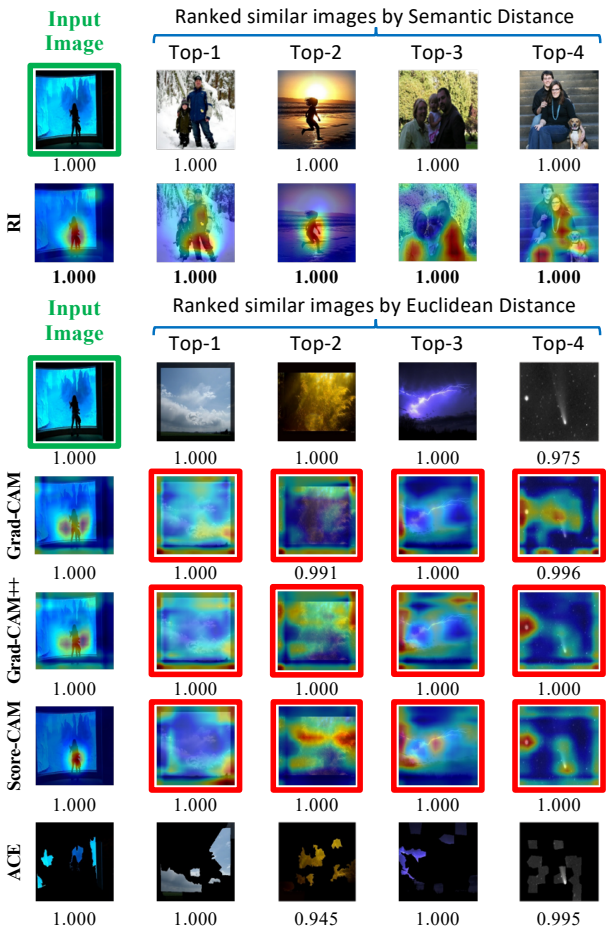


Figure 3. A case study on the FOOD dataset. The first row shows the input image and the similar reference images found by RI. The third row shows the same input image and the similar reference images found by the baseline methods. The other rows are the interpretations of all the methods. The numbers under the images in the first and third row are the prediction scores of the 'non-food' class. The numbers under the interpretations are the prediction scores of 'non-food' on the masked images produced by keeping 20% most important pixels. A higher prediction score means a better interpretation.

As shown in the first row of Figure 3, the input image and the similar reference images found by RI are all predicted as 'non-food' with high prediction scores.

All these images contain human, which indicates the common decision logic to predict these images as 'non-food' is that they contain human bodies instead of food.

We can see from the second row of Figure 3 that this common decision logic is accurately identified by RI. The interpretation computed by RI from the input image correctly identifies the human bodies in the input image and the similar reference images.

These results demonstrate the superior performance of RI in finding representative interpretations that reveal the common decision logic on similar images.

The third row of Figure 3 shows the same input image as the first row, and the similar reference images found by the

baseline methods. These images are perceived as conceptually similar because they all contain large areas of natural scene, thus the common decision logic to predict them as 'non-food' should be the natural scene.

However, we can see from the 4th, 5th and 6th rows that the interpretations produced by Grad-CAM, Grad-CAM++ and Score-CAM all identify the human bodies instead of the natural scene in the input image, but the interpretations on the similar reference images all identify natural scene.

The inconsistency between the interpretations on the input image and the similar reference images demonstrate that Grad-CAM, Grad-CAM++ and Score-CAM cannot produce good representative interpretations to reveal the common decision logic of a CNN in making predictions on a large number of similar images.

We can also see from the last row of Figure 3 that ACE did a fairly good job in consistently identifying the image patches of natural scene in the input image and the reference images. However, as demonstrated by the quantitative experimental results in Section 6.4 and Appendix C, the interpretation quality of ACE is significantly lower than RI due to the high sensitivity of image segmentation and clustering.

C. Representativeness on Reference Images

In this section, we use the reference dataset to quantitatively evaluate how well can an interpretation generated on an input image be reused to interpret predictions made on similar reference images.

We randomly sample 1,000 images from the reference dataset as the input images to generate 1,000 interpretations for each method. For each interpretation, we use AD (7) and AI (8) again for evaluation, except this time we take S to be the 1,000 most similar reference images.

For each of RI, Grad-CAM, Grad-CAM++ and Score-CAM, we use every input image to generate an interpretation. Since ACE requires a set of images to produce one interpretation, we first use an input image to find 49 nearest reference images to the input image in the space of Ω. Then, we use the 50 images including the input image and the 49 nearest images as the input to ACE to generate one interpretation for the input image.

For each of the baseline methods, to evaluate the representativeness of an interpretation computed from an input image x, we first find the top-K nearest reference images to x in the space of Ω using Euclidean distance, and then use these images as the set of similar images S to compute the AD and AI of the interpretation.

For RI, we use the set of top-K similar images ranked by semantic distance to compute the AD and AI of each interpretation.

Figure 4 shows the mean Average Drop (mAD) and the mean Average Increase (mAI) of the 1,000 interpretations generated by each of the methods for different values of K.

Here, $K = 1$ means the set of similar images S contains only the input image x, because x is the most similar to itself. In this case, the mAD and mAI performance are evaluated on the input images only.

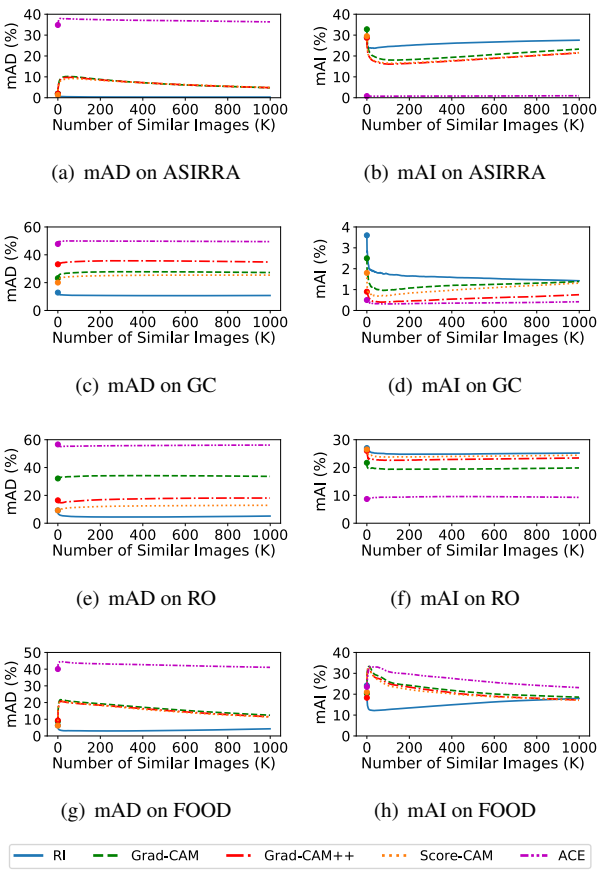


Figure 4. The mean Average Drop (mAD) and mean Average Increase (mAI) performance of all the methods.

We can see that RI achieves the best mAD performance on all the datasets, and it also achieves the best mAI performance on most of the datasets. These results demonstrate the superior performance of RI in producing representative interpretations.

The mAI performance of RI is worse than the other methods in Figure 4(h), because a large proportion of the similar images found by RI on the FOOD dataset has a high prediction score close to 100%. Thus, it is very difficult to further increase the score by masking the images based on interpretations.

Actually, due to the high quality of the representative interpretations produced by RI, most of the masked images produced by RI only have a slight drop of prediction scores. Therefore, as shown in Figure 4(g), RI still achieves a much better mAD performance than all the baseline methods on the FOOD dataset.

We can also see in Figures 4(a), 4(e) and 4(g) that the mAD of Score-CAM is slightly better than RI when $K = 1$. This is because Score-CAM focuses on maximizing the prediction scores of the masked input image, thus it achieves a better mAD on the input image. However, RI focuses on finding the most representative interpretation for both the input image and a large number of similar images. There-

Datasets		ACU-GT of RI (%)	ACU-GT of F (%)	ACU-MD of RI (%)	ACU-MD of F (%)	CVR of RI (%)	CVR of F (%)
ASIRRA	Ref.	99.85 ± 0.04	100.00	99.85 ± 0.04	100.00	99.83 ± 0.05	100.00
	Unseen	98.38 ± 0.12	98.80	99.28 ± 0.08	100.00	99.84 ± 0.14	100.00
GC	Ref.	96.25 ± 0.09	95.34	99.63 ± 0.05	100.00	97.95 ± 0.14	100.00
	Unseen	94.95 ± 0.13	94.00	98.98 ± 0.20	100.00	98.18 ± 0.39	100.00
RO	Ref.	98.63 ± 0.08	99.58	98.98 ± 0.08	100.00	99.69 ± 0.05	100.00
	Unseen	98.14 ± 0.17	97.50	98.44 ± 0.10	100.00	99.88 ± 0.15	100.00
FOOD	Ref.	99.84 ± 0.06	100.00	99.84 ± 0.06	100.00	99.45 ± 0.11	100.00
	Unseen	97.87 ± 0.13	98.50	98.54 ± 0.10	100.00	99.68 ± 0.21	100.00

Table 2. The ACU-GT, ACU-MD and CVR performance of RI and the CNN model F. We run RI for 5 independent times and compute the mean and standard deviation for each of ACU-GT, ACU-MD and CVR. The ACU-GT of F is the prediction accuracy of the CNN model F with respect to the ground truth labels. The ACU-MD of F is always 100% because the prediction results of F are used as the ground truth to compute ACU-MD. The CVR of F is always 100% because F is applicable to predicting the labels of all the images. Ref. and Unseen represents the reference dataset and the unseen dataset, respectively. We set $|V| = 40$ for RO, and $|V| = 20$ for the other datasets.

fore, when $K = 1$, RI achieves a comparable performance with Score-CAM on the input images; and when $K > 1$, RI achieves a much better performance than Score-CAM on the similar images.

D. Prediction Accuracy and Coverage of the Decision Regions Produced by RI

In this section, we evaluate the quality of the interpretations of RI by analyzing the prediction accuracy and the coverage of the decision regions produced by RI.

Recall that each representative interpretation produced by RI for an input image $x \in \mathcal{X}$ is a convex polytope $P(x)$, and $P(x)$ induces a decision region that predicts all the images it covers as $Class(x)$.

If the predictions made by a decision region have a high accuracy, then the corresponding interpretation will be closer to the real decision logic of the CNN F. If the convex polytope of the decision region covers a lot of images, then the interpretation is representative.

Based on the above insight, we design the following experiments to analyze the prediction accuracy and the coverage of the decision regions produced by RI.

For each dataset, we first follow the steps in Section 6.4 to generate a set V of interpretations using RI. Each interpretation $v \in V$ corresponds to a convex polytope that induces a decision region. This gives us a number of $|V|$ decision regions in total.

As illustrated in Section 6.4, we set $|V| = 40$ for RO, and $|V| = 20$ for the other datasets.

For each decision region produced by RI from an input image x, we predict all the images covered by the corresponding convex polytope as $Class(x)$.

If an image $x' \in \mathcal{X}$ is covered by multiple convex polytopes, we predict the label of x' by the decision region induced by the convex polytope that covers the largest number of reference images.

We evaluate the accuracy of the predictions made by the decision regions of RI by the following two types of prediction accuracies.

The first one, denoted by ACU-GT, is the prediction accuracy computed using the ground truth labels of the images. A higher ACU-GT means the decision regions work better in accurately predicting the ground truth labels of the covered images, which further indicates the corresponding interpretations are likely to capture some useful patterns for making accurate predictions.

The second one, denoted by ACU-MD, is the prediction accuracy computed by treating the labels predicted by the CNN model F as ground truth. A higher ACU-MD means the predictions made by the decision regions align better with the predictions made by F, which further indicates the corresponding interpretations are closer to the decision logic of F.

We also evaluate the coverage (CVR) of the decision regions by the proportion of images that are covered by at least one of the $|V|$ decision regions. A larger CVR means a higher representativeness of the interpretations.

We run RI for 5 independent times and compute the mean and standard deviation for each of ACU-GT, ACU-MD and CVR.

Table 2 shows the ACU-GT, ACU-MD and CVR performance of RI and the CNN model F on the reference datasets and unseen datasets of ASIRRA, GC, RO and FOOD.

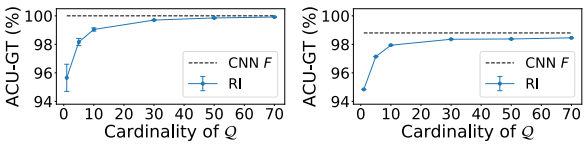
Since the interpretations produced by the baseline methods cannot be used as classifiers to make predictions on images, we cannot report their ACU-GT, ACU-MD and CVR performance.

As shown in Table 2, the decision regions produced by RI achieve very high ACU-GT on both the reference dataset and the unseen dataset. This means the decision regions capture some useful representative patterns to make accurate predictions.

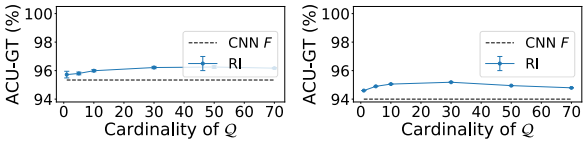
The ACU-GT of RI sometimes outperforms the original CNN model F. This is because the ACU-GT of RI is computed on the covered images, but the ACU-GT of F is computed on the complete set of images, which may contain more misclassified images.

The decision regions also achieve very high ACU-MD on all the datasets. This indicates that the predictions made by the decision regions align very well with the CNN model F. Thus, the corresponding interpretations are closer to the decision logic of F.

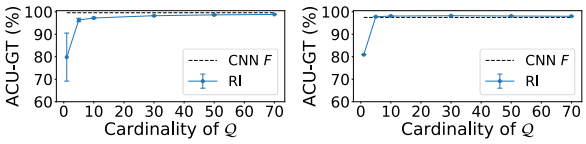
The CVR of RI is also very large. With 40 interpretations for RO and 20 interpretations for the other datasets,



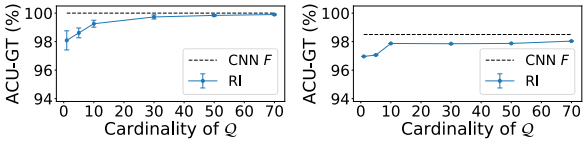
(a) Reference dataset of ASIRRA (b) Unseen dataset of ASIRRA



(c) Reference dataset of GC (d) Unseen dataset of GC

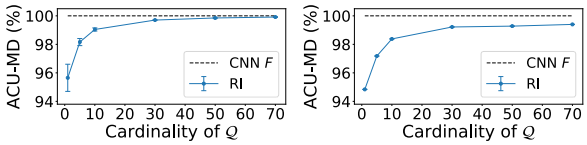


(e) Reference dataset of RO (f) Unseen dataset of RO

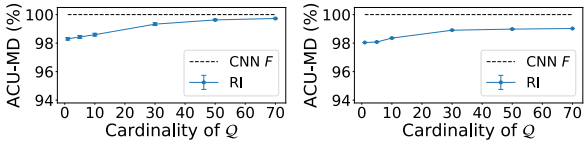


(g) Reference dataset of FOOD (h) Unseen dataset of FOOD

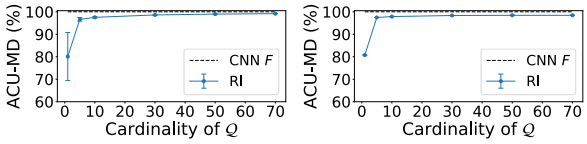
Figure 5. The ACU-GT performance of RI on the reference and unseen datasets of ASIRRA, GC, RO and FOOD.



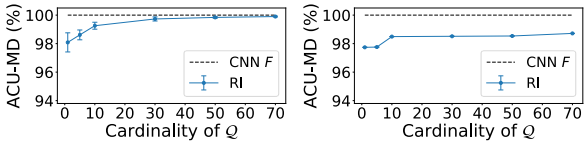
(a) Reference dataset of ASIRRA (b) Unseen dataset of ASIRRA



(c) Reference dataset of GC (d) Unseen dataset of GC



(e) Reference dataset of RO (f) Unseen dataset of RO



(g) Reference dataset of FOOD (h) Unseen dataset of FOOD

Figure 6. The ACU-MD performance on the reference and unseen datasets of ASIRRA, GC, RO and FOOD.

the proportion of images covered by the corresponding decision regions is more than 97% on all the datasets.

Recall that an interpretation of RI generally applies to all the image covered by the corresponding decision region, a large CVR demonstrates the outstanding representativeness of the interpretations produced by RI.

E. Parameter Analysis

In this section, we analyze how the cardinality of $\mathcal{Q}$, denoted by $|\mathcal{Q}|$, affects the ACU-GT and ACU-MD performance of RI on every dataset.

We follow the same experiment setting as Appendix D to compute the mean and standard deviation of the ACU-GT and ACU-MD of RI for 6 different values of $|\mathcal{Q}|$, such as 1, 5, 10, 30, 50 and 70.

Figure 5 shows the ACU-GT of RI and the CNN model F on the reference and unseen datasets of ASIRRA, GC, RO and FOOD.

Each solid point on the blue solid curve shows the mean ACU-GT of RI, and the corresponding error bar shows the standard deviation of the ACU-GT of RI.

The ACU-GT of F is a single scalar for each value of $|\mathcal{Q}|$, thus we draw the ACU-GT of F as a dashed curve without error bars.

Figure 6 is drawn in a similar way as Figure 5 to show the ACU-MD of RI and F on each dataset. The ACU-MD of F is always 100% because the prediction results of F is used as ground truth to compute ACU-MD.

We can see from Figure 5 and Figure 6 that the ACU-GT and ACU-MD of RI are small when $|\mathcal{Q}|$ is small. This is because every convex polytope produced by RI consists of some linear boundaries in $\mathcal{Q}$. If the number of linear boundaries in $\mathcal{Q}$ is too small, a convex polytope produced by RI will not have enough linear boundaries to approximate the complex decision logic of F. Therefore, the ACU-GT and ACU-MD of RI will be compromised.

We can also see that, when the cardinality of $\mathcal{Q}$ increases, the ACU-GT and ACU-MD of RI first increase and then become stable when $|\mathcal{Q}|$ is large. The reason is that increasing the number of linear boundaries in $\mathcal{Q}$ largely improves the descriptive power of the convex polytopes produced by RI, thus the ACU-GT and ACU-MD of RI increases significantly in the beginning. However, when the number of linear boundaries in $\mathcal{Q}$ is large, existing linear boundaries in $\mathcal{Q}$ are good enough to produce high-quality convex polytopes, thus adding newly sampled linear boundaries into $\mathcal{Q}$ will not further increase the ACU-GT and ACU-MD of RI very much. In consequence, the ACU-GT and ACU-MD of RI become stable when $\mathcal{Q}$ is large.

Recall that a high ACU-GT indicates that the interpretations produced by RI capture some useful patterns of data to make accurate predictions; and a high ACU-MD means the

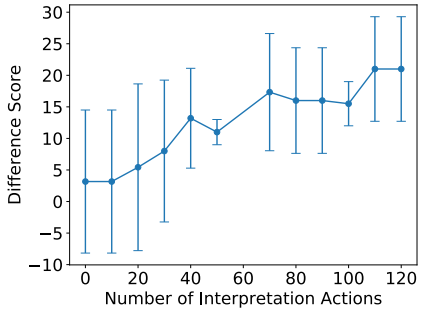


Figure 7. The relationship between the difference score and the number of interpretation actions. The x-axis represents the number of interpretation actions, which is between 0 and 150. The y-axis represents the difference score.

interpretations produced by RI is close to the decision logic of F. According to the results in Figure 5 and Figure 6, since the ACU-GT and ACU-MD of RI are large and stable on all the datasets when $|\mathcal{Q}|$ is larger than 50, we simply set $|\mathcal{Q}| = 50$ for RI to achieve outstanding interpretation performance in our experiments.

F. A/B Test on Retina OCT Dataset

In this section, we conduct an A/B test on the Retina OCT (RO) dataset to demonstrate the effectiveness of RI in improving the diagnosis accuracy of human on retina disease.

The setting of the A/B test is as follows.

Task: we formulate a binary classification task using images from the two classes of NORMAL and DME in the RO dataset. NORMAL represents the images of normal retina and DME is a type of retina disease. The goal of the task is to predict whether an input image is NORMAL or DME.

The CNN model and the interpretation method: we train a VGG-19 model [53] to achieve a testing accuracy of 97% for the binary classification task between NORMAL and DME. We use RI to produce representative interpretations on the VGG-19 model.

Subjects: the subjects of the A/B test is a group of 20 people without any knowledge background on retina disease. To prepare these people for the binary classification task, we give the same written tutorial to each subject to teach them the basic skills to distinguish between NORMAL and DME. Every subject has 10 minutes to read the tutorial.

Test A: the test A consists of 50 multiple-choices questions. Each question requires the subject to answer whether a retina image is NORMAL or DME. The choices of answers are: 'Definite DME', 'Maybe DME', 'Not sure', 'Maybe NORMAL', and 'Definite NORMAL'. A 'Not sure' answer receives a score of 0. For answers with 'Maybe', a correct one receives a score of +1, but a wrong one receives a score of -1. For answers with 'Definite', a correct one receives a score of +2, but a wrong one receives a score of -2. Every subject taking test A will see the prediction results of the CNN model on the input retina images. However, to simulate the practical scenario that many people may not trust a machine learning model in making diagnosis, we lie to every subject that the CNN model only has a testing accuracy of 80%. Every subject is taking the same set of 50 questions, but the order of the questions is randomly generated for each subject. The sum of the scores of the 50 questions are collected as the final score for each subject in test A.

Test B: the test B follows exactly the setting of test A. The only difference is that a subject is able to see the interpretation result generated by RI for each input retina image. We allow a subject to freely choose whether or not to see the interpretation result for the input image of each question. We log three types of **interpretation actions** when a subject sees the interpretation result; these actions include: (i) 'show similar images' shows the similar retina images to the input image ranked by RI; (ii) 'show heat map' shows the heat map generated by RI on the input retina image and the similar retina images; and (iii) 'Zoom in' enlarges the similar images to show more details. The sum of the scores of the 50 questions are collected as the final score for each subject in test B.

Every subject is required to take Test A first and then take test B. Since the orders of the 50 questions are randomly generated for every test, it is very difficult for a subject to memorize his/her answers in test A when taking test B.

Every subject produces a final score for test A, denoted by 'Score A', and a final score for test B, denoted by 'Score B'. We collect the **difference score** between Score B and Score A, that is, Score B minus Score A, for each subject. This produces 20 difference scores.

A larger difference score means Score B is higher than Score A, which indicates using interpretations produced by RI can improve the diagnosis accuracy of human on retina disease.

Recall that we also log the interpretation actions of each subject in test B. We collect the total number of interpretation actions of each subject to measure how often a subject uses the interpretation results produced by RI.

We draw the results produced by the 20 subjects as the blue points and the corresponding error bars in Figure 7. Denote by (x, y) the coordinates of a blue point, y is the mean of the difference scores of all the subjects whose numbers of actions fall into the interval of $(x, x + 30]$. The corresponding error bar shows the standard deviation of the difference scores of the subjects. There is no blue point for $x = 60$ since there is no participant whose number of interpretation actions is between 60 and 90.

Figure 5 shows that a more frequent use of the interpretation results produced by RI contributes to a higher difference score. This demonstrates the high effectiveness of RI in helping people making more accurate diagnosis on retina images.

G. More Interpretation Examples of RI

We present more examples in Figure 8 to show the good interpretation performance of RI on each of ASIRRA, GC, RO and FOOD.

We can see from the results that the representative interpretations produced by RI always highlight meaningful

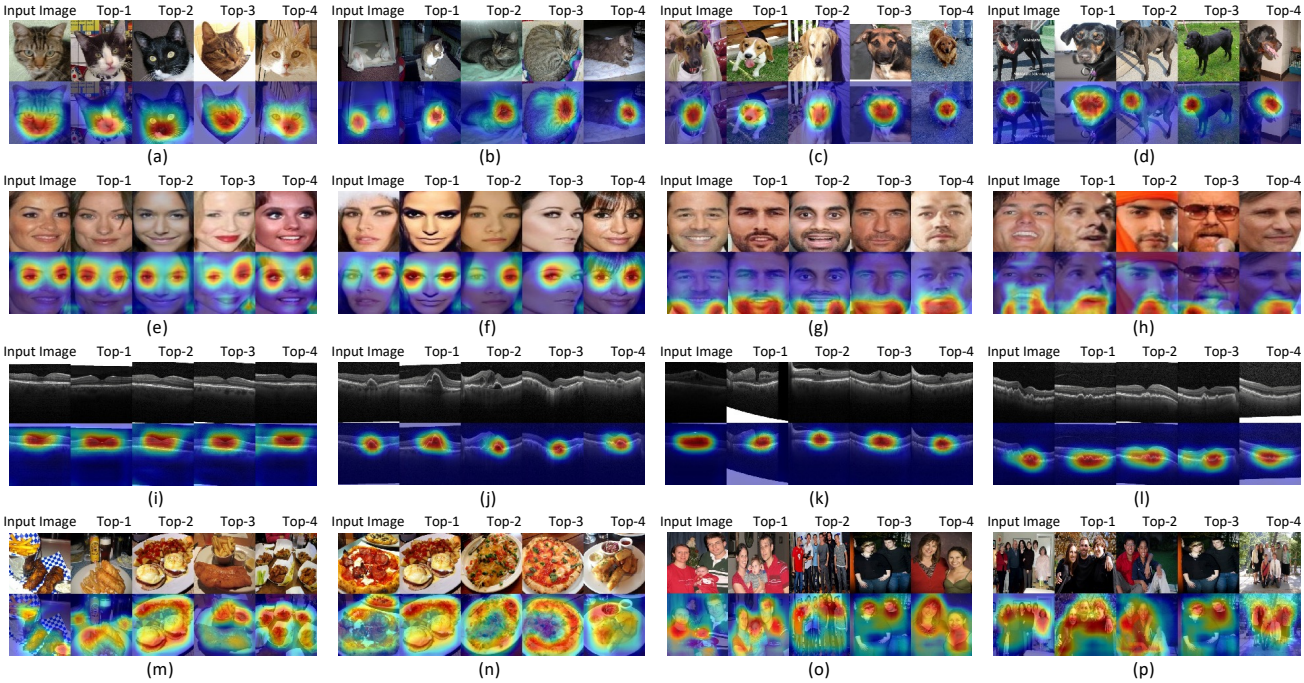


Figure 8. Representative interpretations produced by RI on ASIRRA, GC, RO and FOOD. For each figure, the first column is the input image and the rest of the columns are the similar images ranked by RI. (a) and (b) show the interpretation results on two images of the class 'cat' in ASIRRA. (c) and (d) show the interpretation results on two images of the class 'dog' in ASIRRA. (e) and (f) show the interpretation results on two images of the class 'female' in GC. (g) and (h) show the interpretation results on two images of the class 'male' in GC. (i), (j), (k) and (l) show the interpretation results on the images of the classes 'NORMAL', 'CNV', 'DME' and 'DRUSEN' in RO, respectively. (m) and (n) show the interpretation results on two images of the class 'food' in FOOD. (o) and (p) show the interpretation results on two images of the class 'non-food' in FOOD.

common parts of the input image and the similar images. For example, the faces of cats and dogs in Figures 8(a)-8(d), and the beard of male in Figures 8(g) and 8(h).

Obviously, showing similar images with common highlighted parts as the input image makes our interpretations more convincing than showing only the interpretation on the input image.

The results in Figure 8 further demonstrate the outstanding performance of RI in producing representative interpretations to reveal the common decision logic of a CNN on an input image as well as the images that are similar to the input image.