Article

Towards a Reliable Evaluation of Local Interpretation Methods

Jun Li ^{1,2,3,4}, Daoyu Lin ^{1,2}, Yang Wang ^{1,2}, Guangluan Xu ^{1,2} and Chibiao Ding ^{1,3,4,*}

- ¹ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; lijun215@mailsucas.ac.cn (J.L.); lindy@aircas.ac.cn (D.L.); ywang1@mail.ie.ac.cn (Y.W.); gluanxu@mail.ie.ac.cn (G.X.)
- ² Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
- ³ University of Chinese Academy of Sciences, Beijing 100094, China
- ⁴ School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
- * Correspondence: cbding@mail.ie.ac.cn

Abstract: The growing use of deep neural networks in critical applications is making interpretability urgently to be solved. Local interpretation methods are the most prevalent and accepted approach for understanding and interpreting deep neural networks. How to effectively evaluate the local interpretation methods is challenging. To address this question, a unified evaluation framework is proposed, which assesses local interpretation methods from three dimensions: accuracy, persuasibility and class discriminativeness. Specifically, in order to assess correctness, we designed an interactive user feature annotation tool to provide ground truth for local interpretation methods. To verify the usefulness of the interpretation method, we iteratively display part of the interpretation results, and then ask users whether they agree with the category information. At the same time, we designed and built a set of evaluation data sets with a rich hierarchical structure. Surprisingly, one finding is that the existing visual interpretation methods cannot satisfy all evaluation dimensions at the same time, and each has its own shortcomings.

Keywords: deep learning interpretability evaluation; prediction explanation; saliency map

Citation: Li, J.; Lin, D.; Wang, Y.; Xu, G.; Ding, C. Towards a Reliable Evaluation of Local Interpretation Methods. *Appl. Sci.* **2021**, *11*, 2732. <https://doi.org/10.3390/app11062732>

Academic Editor: Ángel González-Prieto

Received: 25 January 2021
Accepted: 16 March 2021
Published: 18 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Machine learning technology that automatically discovers patterns and structures in massive amounts of data is becoming more and more widely used in real life [1–3]. In particular, deep convolutional neural networks (DCNNs) have demonstrated impressive results in many real scenarios, such as image classification [4,5], speech recognition [6], and recommendation systems [7]. Despite the success of deep learning methods in these tasks, the opacity of this method has made it criticized [8]. The main reason is that the internal working mechanism is not clear because of the large number of model parameters and the complex structure of the model. Due to the lack of understanding of the internal working mechanism, DCNNs are often regarded as “black box” models, which leads to the decision results of DCNNs being not convincing to humans. The need for interpretation of DCNNs arises in this situation.

Providing human-intelligible explanations for the decisions of DCNNs is the ultimate goal of Explainable Artificial Intelligence (XAI) [9,10]. Existing work on interpretability mainly involves training interpretable machine learning [11,12] and explaining black-box models [13–15]. Here, we mainly focus on the methods of explaining the black box model. Feature-based post-hoc explanatory methods [9] are the most prevalent and are the natural way to interpret neural networks. This type of visual interpretation method aims to generate a feature saliency map (FSM) corresponding to the input image from the prediction results of the model, namely, the generated feature saliency map quantifies how much each input feature contributes to the model output.

In line with this research, although these approaches explain the predictive behavior of DCNNs to a certain extent, these interpretation methods give inconsistent visual explanations for the same prediction result, thus it is not clear to the user which interpretation method is reliable. For example, as shown in Figure 1, for the same trained model and the same prediction result (Figure 1A), the feature explanations given by the three algorithms (Gradient-Weighted Class Activation Mapping—Grad-cam [16], Mask [17], Rise [18]) are the elephant's face and nose, the elephant's trunk and nose and the elephant's face and tusk respectively (Figure 1B). Therefore, how to verify the faithfulness of these methods is a problem worthy of attention. Especially when users are faced with these different interpretations and they do not know which interpretation results to trust. It motivates us to evaluate which of these interpretation methods is more human-friendly.

Figure 1. (A) shows the decision-making process of the trained model. (B) shows the results of different interpretation methods for the same model and the same prediction result (From left to right: Grad-cam, Mask, Rise). Different interpretation algorithms give different reasons for prediction, which explanation should users trust and which one is more convincing?

The evaluation of interpretation methods is a challenging problem. Specifically, there are two main challenges. One is that the correctness of the FSM cannot be measured due to the lack of ground truth (we do not know which input features are in fact important to a model). The correctness of the FSM refers to correctly reflecting the contribution of each feature to the model prediction results. For example, some study [19] has shown that certain visual interpretation methods are misleading to users, and some interpretation methods are quite easy to be fooled by adversarial examples [20]. In order to solve the problem of no ground truth, we construct an evaluation data set and provide the ground truth of dataset.

Another challenge is that there is no unified metric for visual interpretation methods. Currently the three types of metrics proposed by Finale and Been [21] are accepted, application-based metric [22], function-based metric [23], and human-based metric [24]. Application-grounded metric mainly refers to conducting human experiments within real application scenarios, such as doctors performing diagnoses. Functionally-grounded metric does not require human experiments and completes the evaluation process with the help of proxy tasks. Human-based metric mainly means that the explanation should be

friendly to humans, because no matter how the explanation method is designed, if the explanation is unreasonable or meaningless to the user, then this explanation method loses its value. Note that in this paper we mainly focus on human-based metric. In order to evaluate human satisfaction with different interpretation methods, we design an “object recognition game” to measure the degree of user satisfaction with different interpretation methods by occluding the saliency maps.

In this paper, we first define three attributes that a human-friendly explanation should satisfy: high accuracy, high persuasibility and high class-discriminativeness. We then manually construct a set of datasets so as to comprehensively evaluate the performance of different interpretation methods. In order to bridge the gap without ground-truth, we use the idea of human annotation to solve the problem. First, the input samples are preprocessed by super-pixel segmentation, and then the user is allowed to select important feature blocks. To the best of our knowledge, we are the first to calibrate key feature blocks in the pixel space of the input image by manual annotation method. For persuasive performance, we design an interesting object recognition game to evaluate different interpretation methods by the user study method. In addition, we use point games to measure the class-discriminativeness of interpretation methods. To summarize, the main contributions to this paper are:

- We innovatively constructed an evaluation data set with three branches to comprehensively measure the performance of different visual interpretation methods. The dataset and human-grounded benchmark are available at this link: <https://github.com/muzi-8/Reliable-Evaluation-of-XAI>, accessed on 10 March 2021.
- Using the manual feature annotation method, we measure the correctness of the visual interpretation method through the important feature blocks obtained.
- Through carefully designed object recognition games, we effectively measured the persuasiveness of the three visual interpretation methods.

The reminder of this paper is organized as follows. The two lines related works are introduced in Section 2. The proposed evaluation method is presented in Section 3. Section 4 demonstrates implementation details and experimental results. Section 5 discusses and analyzes the experiment results. The conclusion is drawn in Section 6.

2. Related Work

There are lots of literatures [25–31] on the research of feature-based post-hoc explanatory methods. Some review articles can be found in [32–38]. Existing research work can be roughly summarized as attribution method and evaluation method.

2.1. Attribution Method

Attribution method is the classical feature-level interpretation method to explore the connection between input and output via identifying features importance with respect to model prediction. The final result of this interpretation method is to generate a FSM highlighting the importance of each input feature to the prediction. The gradient-based contribution method mainly uses gradient information to obtain the importance of input features. Three different strategies have been proposed to capture input features relative importance. They are backpropagation-based method [27,28], activation-based method [16,29] and perturbation-based method [17,18,30]. Simonyan et al. [27] assumes the most important feature is the input pixel that makes the output the most sensitive, and quantifies the importance of each input via the pixel gradient itself. Guided prop [28] modifies the backpropagation condition and only allows the positive gradient to propagate back to each neuron. CAM [29] and Grad-cam [16] leverage high-level information of deep learning model and weight sum of the last convolutional feature maps. Perturbation-based attribution method is another technique for explaining predictions of deep learning model. These methods perturbate the input with some certain reference value and find the smallest mask through an optimized way to maximize the output result. Fong et al. [17] optimized the method to make the generated mask smooth and meaningful. Petsiuk et al. [18] ran-

domly performs binarization sampling on the input image, and finally obtains the saliency map by weighting the predicted value. Qi et al. [30] improved the optimization process with integrated gradient.

2.2. Evaluation Method

Due to the lack of ground truth, we do not know which pixel is in fact important to a model. Existing evaluation methods can be classified into three categories, namely, removing pixel features [17,18,39], setting relative ground truth [20,26,40] and user-oriented measurement [41,42].

2.2.1. Removing Pixel Feature

A class of intuitive methods is to directly remove important features, and then perform forward inference. Finally, the correctness of the interpretation method is measured according to the change in model accuracy. These methods are also distinguished by whether or not to retrain the model. Work [17,18] is to observe the change of the class probability on the original trained network. However, Sara et al. [39] removes a fraction of important input features according to each interpretation method, and retrain the model from random initialization with these modified datasets, which points out that training a new model is critical.

2.2.2. Setting Relative Ground Truth

Another different method is to set ground truth from different perspectives. Akshay-varun et al. [20] introduces adversarial patches as a true cause of prediction and shows that Grad-cam interpretation method is unreliability and easily be fooled by the adversarial example. Mengjiao et al. [26] constructs carefully semi-natural dataset by pasting object pixels into scene image and trains models with these dataset. Meanwhile, they suggest that certain interpretation methods can produce false positive explanations for users. However, the shortcomings are that they do not focus on how humans understand explanations. Oramas et al. [40] constructs a synthetic dataset to evaluate the correctness of feature importance ranking.

2.2.3. User-Oriented Measurement

Human annotation [41] and human study [42] are two classical ways to measure visual interpretation feature representation. In some classic tasks, human annotation methods are commonly used, such as object detection and semantic segmentation tasks, and evaluating the performance of the model with the help of these labeled ground truth. Zhou et al. [41] uses crowdsourcing to annotate the visual concept of the scene, and intuitively explains the deep visual representation by measuring the overlap between the image area activated by the neuron and the annotation. The reason why the human study strategy is used is that it can flexibly set experimental requirements and verification objectives. Linsley et al. [42] introduces a web-based clicktationary game to measure the relationship between human observers and FSM.

3. Human-Friendly Evaluation

In this section, we first introduce the background knowledge of the evaluation method, and then introduce our data set construction process and an overview of our unified evaluation framework. Finally, the evaluation process of the three dimensions of accuracy, persuasibility and class-discriminateness are introduced respectively.

3.1. Preliminaries

Gradient-Weighted Class Activation Mapping (Grad-cam) is the most popular interpretation method that uses a gradient-weighted linear combination of activation from convolutional layers. This approach takes the internal network state into account and produces a coarse FSM that is most responsible for model decision.

First of all, the importance of k -th channel feature map for c -th class α_k^c can be calculated based on the last fully-connected layer of fine-tuned convolutional neural network (CNN) module. After, the feature maps on the last convolutional layer of fine-tuned CNN models are combined with different weights and Rectified Linear Unit (RELU) function so as to eliminate all negative values in $\text{Grad-cam} = \text{Relu}(\sum_k \alpha_k^c A^k)$.

Meaningful image perturbations (Mask) is a perturbation-based interpretation method that perturbs the input with optimization to estimate the importance of each input pixel. This method is model-agnostic and testable because of explicit and interpretable image perturbations. The perturbed image Mask is defined by $\text{Mask} = m \odot x + (1 - m) \odot r$, where m is a mask, x is the input image, and r is a reference image. $\odot$ denotes element-wise multiplication.

Randomized input sampling (Rise) is a black-box-oriented interpretation method. The main idea behind this approach is to sample the input via random masks and then monitor the model class probability. Rise generates the final visual interpretation by a linear combination of the random binary masks. This method is defined by $\text{Rise} = \sum_m f(x \odot M) \odot M$, where M is a random binary mask with distribution D . $\odot$ denotes element-wise multiplication.

3.2. Framework of Evaluation

Our evaluation framework consists of model decision process (a), generating interpretation process (b) and user evaluation process (c) (Figure 2). In the decision-making stage, we adopt a pre-trained model (Vgg-16). It is worth noting that we pay more attention to the model prediction results. At the interpretation stage, we investigate different interpretation methods. Specifically, we constructed a complete set of evaluation data sets at this stage. In the process of user evaluation, we focus on three dimensions of explanations: accuracy, persuasibility and class-discriminateness. Study [26] has shown that visual interpretation methods are making mistakes. For example, methods assign high attributions to unimportant features (False Positive/FP) and low attributions to important features (False Negative/FN). Accuracy refers to the explanation should assign high attributions to important features (True Positive/TP) and low attributions to unimportant features (True Negative/TN). Persuasibility means that the user can accurately understand the explanation and the explanation result is meaningful. Class-discriminateness represents the explanation only be given on the relevant objects and not cause user ambiguity. These three dimensions might reflect essentially what a good explanation is.

Figure 2. Overview of our evaluation framework, which contains three parts model decision process (a), generating interpretation process (b) and user-oriented evaluation process (c) (Note that the blue box represents the existing research results, the orange box is our innovative work).

Definition 1: We define the accuracy of explanation as accurately highlighting image features that play a key role in model decision making.

Definition 2: We define the persuasibility of explanation as the satisfaction degree to human users, aiming to measure the subjective satisfaction of explanation.

Definition 3: We define the class-discriminateness as whether the explanation method has the ability to distinguish different object targets.

3.3. Evaluation Dataset Construction

All the evaluation data sets come from ImageNet [43]. The constructed dataset can be divided into three branches according to the category and the number of objects contained in the image. As shown in Figure 3, the three branches are images containing only a single object (DataS), images containing multiple objects of the same category (DataM), and images containing two objects of different categories (DataT).

- DataS can be considered into three main branches—vehicles, animals, and clothing. The vehicle category contains four fine-grained sub-classes (e.g., police cars, ambulances, school buses, tanks). Animal category consists of five fine-grained sub-classes (e.g., goat, cat, panda, elephant, polar bear). The clothing branch contains a category of military uniforms. These fine-grained sub-classes have their unique discriminative feature. For example, the unique feature of the police car is the police text logo printed on the body, and Ivory only appears in the elephant category. This branch consists of a total of 40 test images, with four images in each sub-class.
- DataM consists of 10 categories (e.g., cheetah, dumbbells, peppers). Each image in this category contains at least two objects. Each sub-class contains one test example respectively.
- DataT means that two categories appear in one image at the same time. Each image contains two categories that belong to the original 1000 categories. From the ImageNet dataset, we collected five test samples. The reason why these data exist in ImageNet is that researchers ignore the inherent multi-label attributes of the image when letting users label the data.

Figure 3. Schematic diagram of the hierarchical structure of the evaluation dataset. The superscript number indicates the number of each type of image.

3.4. Overview of Accuracy Evaluation

The purpose of Accuracy Evaluation (AE) is to quantitatively evaluate different interpretation methods through the ground truth of features labeled by users. In order to realize this idea, we designed a web-based feature annotation tool to interactively acquire key features of different categories of images. Figure 4 shows the whole process of AE, which consists of four parts: evaluation data, feature segmentation, interactive annotation tool design and quantitative evaluation. The following details our interactive feature annotation core module(Figure 4c).

Figure 4. Accuracy Evaluation (AE) pipeline of interpretation methods. There are four steps in total: dataset (DataS), pixel segmentation, pixel feature selection and evaluation process.

Figure 5 shows our annotation interface and the process of generating ground truth maps. We created ground truth map through a two-step procedure. First, the feature block for every image selected by each participant is recorded (Figure 5A). In order to facilitate the user to select features, inspired by [44], we used the SLIC algorithm [45] to perform superpixel segmentation preprocessing operations on the evaluation dataset. Second, ground truth map is generated by integrating the feature blocks selected by the participants (Figure 5B). (Note that in order to make the ground truth map more accurate while taking into account operability, we recruited a total of 40 high-quality participants).

Figure 5. Overview of feature annotation interface. (A) is our main interface, which displays the segmented image and category information to be labeled, and also informs the user of the number of remaining labeled images. (B) means that we obtain the final ground truth by counting the feature blocks marked by different users.

In order to ensure the quality of the feature block selected by the user, we have carried out a series of user instructions before the user operation. For example, we informed the user to select the most important feature blocks as much as possible, at least 4 feature blocks, and no more than 8 for each image (See Appendix B for the detailed specific user instructions). At the same time, our annotation tool can also automatically monitor whether the number of feature blocks marked by the user meets 4–8, and the next image can be labeled only when the number of selected feature blocks is satisfied. Let $U_1(I), U_2(I), \dots, U_m(I)$ denotes the set of feature blocks annotated by $user_1, user_2, \dots, user_m$, respectively. Here, each $U(I) \subset Seg(I)$, which contains a total of 100 feature blocks in image I . m represents the number of users. Every $U_j(I)$ is a binary mask image. When the user selects a pixel block, all the pixel values of the pixel block are 1, otherwise the pixel value of the pixel block is 0. Finally, we count the results of all user annotations, in other words, we sum the pixel values of the corresponding positions of all user annotation results of the same image. Therefore, the maximum value of pixel value is 40 (this means that the pixel blocks

here are marked by all users) and the minimum value is 0. The final ground truth (GT) is normalized to 0–1. The formalization is as follows.

$$GT = \frac{1}{m} \sum_{j=1}^m U_j(I) \quad (1)$$

According to the obtained ground truth map of the category feature block, we quantitatively calculate the correctness of different interpretation algorithms. The formal definition of the Mean Absolute Error (MAE) is as follows.

$$MAE = \frac{1}{c} \sum_{k=1}^n \sum_{(i,j) \in I} |E(\hat{y})_{k,(i,j)} - GT_{k,(i,j)}| \quad (2)$$

where n is the number of samples (n is equal to 30 in this experiment), k denotes a specific image among the 30 images, and the indices (i, j) run over all the pixels of the figure and the subindex (i, j) denotes the importance value of that pixel. The c is the constant value $50176 = 224 \cdot 224$, the total number of pixels of the image. $E(\hat{y})$ denotes the input feature importance distribution obtained by different interpretation methods.

3.5. Overview of Persuasibility Evaluation

The primary goal of persuasibility evaluation (PE) is to quantitatively measure the degree of user satisfaction with different interpretation methods. We designed a game that occludes partial explanations and allowed users to interactively answer object name to achieve this goal. The main components are shown in Figure 6. PE takes a trained VGG-16 CNN model and the corresponding evaluation dataset (DataS) as the input (Figure 6a). In the data processing stage, FSM about the prediction results are generated and feature occlusion operations on these interpretation images are performed (Figure 6b). The occlude images is then passed to the recognition object module. This module is developed in the form of questionnaires, including three rounds to collect user results interactively (Figure 6c). Finally, we evaluate different interpretation methods by analyzing user feedback data. Specifically, we use box plots [46] to analyze user satisfaction ratings, and use Sankey Diagram [47] to demonstrate user selection changes (Figure 6d).

Figure 6. PE pipeline of interpretation methods. There are four steps in total: dataset (DataS), image occlusion, object recognition process and evaluation process.

For the entire satisfaction evaluation process, user interactively recognition module is the core part and the occlusion of the FSM directly affects the user's difficulty in identifying objects and the evaluation results. We assume that a good interpretation method should be able to accurately find the most important features, therefore, the range of feature areas in the occlusion map should also be as small as possible. Based on the interpretation results, we empirically set two different thresholds: $\lambda_1 = 3\%$ (pixel features ranked in the top 3%) and $\lambda_2 = 8\%$ (pixel features ranked in the top 8%). Take the elephant category as an example (Figure 7), for different interpretation methods, the occlusion images of these two thresholds can reflect the relevant features of the elephant category to a certain extent, and are acceptable for the evaluation process.

Figure 7. Three visual interpretation algorithms retain 3 % and 8% feature occlusion maps.

Once the recognition game starts, the user will progressively go through three rounds to name the category and score the category satisfaction of occluded image (Figure 8). In round 1, the top-3 occlusion image is first shown to the user, and the participants are asked whether they can recognize the object. If the user answers yes, the specific name of the object needs to be given. Otherwise it will go to round 2. In round 2, we further provided the category information of the corresponding occluded image, and at the same time asked the user whether he agreed with the given category, and scored 1–5. In round 2, users score 1–5 for category satisfaction. If the user scores 1 or 5 points, the entire process ends. If the user chooses 2, 3, 4 points, the round 3 is entered. In round 3, the feature range of the occluded image will be expanded from 3% to 8%, and the user scores 1–5 for the 8% occluded image. To sum up, the PE included an object name recognition question in round 1 and a 5-point Likert-scale [48] question in round 2 and round 3 (See Appendix A for the detailed recognition interface).

Figure 8. Flow chart of user persuasion method with three rounds (round1, round2 and round3), different colors indicate different round processes.

3.6. Overview of Class-Discriminateness Evaluation

The class-discriminateness refers to the ability of visual interpretation algorithms to help distinguish between classes. We assume that a good visual interpretation algorithm

will only highlight features related to the predicted class, but will not highlight image areas of other class. Specifically, when an image contains two kinds of objects, the visual interpretation method should accurately highlight the category features related to the prediction result. For example, the model predicts a test case that contains both bananas and pineapples, and the final judgement result is bananas. The saliency map generated by the visual interpretation algorithm should only highlight features related to bananas, not pineapples.

Pointing Game experiment is first introduced in [49] to evaluate the discriminativeness of different visualization methods for localizing target objects in scenes. Similar to [18], we also use Pointing Game to evaluate the discriminativeness of various visual explanation methods for localizing target objects in the image, which identify whether the prediction is related to the object in the image. We extract the maximum point on the different attention map and we call this the pointing game. Pointing Game asks the CNN model to point at an object of designated category in the image. Meanwhile, the pointing game does not require highlighting the full extent of an object. If the maximum point lies on one of the instances of the object category, it is considered as hitting the object, otherwise a miss is counted. The overall performance of different interpretation method is measured by the average number of hits across all sample. Finally, The pointing game value is given by $\frac{\#Hits}{\#Hits + \#Misses}$. Since the pointing game only objectively counts the maximum pixel position of the attention map and it does not impact the CNN model's classification accuracy, so it can evaluate different types of attention maps more fairly. Because the classification model relies on the target object information, this metric value should be high for a good visual interpretation.

4. Experiments and Results

4.1. Participant Recruitment and Experimental Design

We strictly recruited 105 participants (50 % male 50% and female, age 20–30, basically covering all majors (Appendix C)). To facilitate the verification of experimental hypotheses, we conducted all experiments based on the pre-trained VGG-16 and aimed at image classification tasks. For the accuracy experiment, We selected 5 and 40 participants from 105 participants, and compared the results marked by different numbers, and informed them of the precautions (Appendix B). From all experimental data were randomly selected three test samples from each category of DataS, which was a total of 30 images. For the persuasibility experiment, we designed our own questionnaire powered by www.wjx.cn, accessed on 10 March 2021, which provides functions equivalent to Amazon Mechanical Turk. We divided the participants into three groups, each with 35 end-users. All experimental data came from the DataS (10 remaining samples) and DataM we built, which was a total of 20 images. For the class-discriminativeness experiment, the experimental data came from the entire constructed data set, including DataS, DataM and DataT. In short, the total amount of experimental data was 55 images.

4.2. Evaluation of Accuracy

Figure 9 shows the human attention masks from 40 unique participants, Here we only show 10 image samples from 10 categories of the evaluation data set DataS, the remaining 20 image samples can be seen in Appendices D and E. The ground truth for image datasets (DataS) were stored in the format of grayscale (0–1), which had the same size of original images (224×224). The human attention masks are displayed with the aid of pseudo-color images. The red feature block represents a higher gray value, which means that the user considers the most important, and vice versa. The blue feature block represents a lower gray value, which means that users had the least number of annotations.

Figure 9. Examples of ground truth result of the relevant feature block marked by the user of each category. **(TOP)** The 10 input image samples in 10 categories from the evaluation data set DataS. **(Middle)** Segmented image with 100 feature blocks obtained according to the SLIC segmentation algorithm. **(Bottom)** Resulting multi-participants human attention mask. Each image is annotated by 40 unique participants.

Figure 10 shows the similarity result of 30 different samples with three different interpretation methods relative to the human attention (ground truth) mask. The similarity measurement result indicated the similarity between the FSM obtained by the interpretation method and the human-attention mask. The smaller the value, the more similar. At the same time, we calculated the average value of each interpretation algorithm relative to ground-truth. As shown in the Table 1, Grad-cam interpretation algorithm had the lowest MAE value, which is the best correctness. The correctness of rise interpretation method was ranked second.

Figure 10. The similarity measure result of the feature importance of different interpretation methods relative to the ground truth value on the DataS data (Police car, Ambulance, School Bus, Tank, Cat, Goat, Panda, Polar Bear, Elephant and Uniform).

Table 1. Accuracy results of different interpretation algorithms (the smallest value is the best).

Interpretation Methods	Grad-Cam	Mask	Rise
MAE	0.23	0.43	0.31

We statistically calculated the confusion matrix (Table 2) of different visual interpretation algorithms. In order to quantitatively evaluate different visual interpretation algorithms, We have calculated the F1-score (Table 3) of each different interpretation method.

Table 2. Confusion matrix of different interpretation algorithms (P means Positive, N means Negative).

	Grad-Cam			Mask			Rise	
	P	N		P	N		P	N
P	1.54	1.6	P	1.76	1.4	P	1.5	1.69
N	4	92.7	N	10.19	86.65	N	5.6	91.21

Table 3. F1-score results of different interpretation algorithms. (The larger value is the best).

	Grad-Cam	Mask	Rise
F1-score	0.35	0.23	0.29

We found that when a different evaluation metric was used to measure the correctness of the interpretation method, the results were completely consistent. The correctness of the Grad-cam algorithm was the best, because this type of interpretation algorithm not only considered the change of the prediction result relative to the feature, but also took into account the internal abstract feature representation state of the model.

4.3. Evaluation of Persuasibility

Figure 11 shows the statistical results of the number of users who could answer the object name based on the top 3% feature occlusion map based on different interpretation methods in Round 1. Figure 12 shows the correct number of users who gave the object name in all the different interpretation methods. We can see that the explanation given by the mask algorithm had many users who choose to answer the name of the object, but the correct rate of the answer was very low. For the interpretation result given by the Grad-cam algorithm, the correct rate of the user's answer was the highest.

Figure 11. For different interpretation algorithms under top-3% occlusion, the statistical results of object names that users can identify. Each interpretation method is tested with 10 images from DataS and 35 users.

Figure 12. The statistical result of the correct object name given by the user. Each interpretation method is tested with 10 images from DataS and 35 users.

We further analyzed the user's satisfaction with the interpretation given by different interpretation algorithms in Round 2 and Round 3. Figure 13 shows the distribution of user satisfaction with different explanation methods after informing users of category information. Figure 14 shows the distribution of user satisfaction regarding whether to agree with category information after the feature interpretation areas was expanded from 3% to 8%. We can see that when the user satisfaction was 1, for the Mask algorithm, the number of users was the largest. However, when the user satisfaction is 3, 4, or 5, the Grad-cam algorithm has a larger number of users compared to the Mask and Rise algorithms.

Figure 13. The user satisfaction distribution of different interpretation algorithms under the top-3% occlusion. Each interpretation method is tested with 10 images from DataS and 35 users. Satisfaction score is 1–5 (1 means most dissatisfied, 5 means most satisfied).

Figure 14. The user satisfaction distribution of different interpretation algorithms under the top-8% occlusion. Each interpretation method is tested with 10 images from DataS and 35 users. Satisfaction score is 1–5 (1 means most dissatisfied, 5 means most satisfied).

We used a Sankey Diagram to visually show how users' choices changed after expanding the range of features (Figures 15–17). The first column of the Sankey Diagram indicates the gender of the user, the second column indicates whether the user understands machine learning, the third column indicates the satisfaction distribution for the category in the first round (Note that the data of strongly agree and strongly disagree are not displayed here, because the game ended directly after selecting these two options), and the fourth column indicates the satisfaction score for the category in the second round. By comparing the Sankey Diagram of three interpretation methods, we can intuitively see that the Grad-cam interpretation method could best convince users, the Rise interpretation method was the second, and the Mask interpretation method had the worst effect. Take the disagree option in the third column as an example, for the Grad-cam interpretation method, the user chose disagree for the first time. After increasing the feature range, most users' choices flowed to the agree option. For the rise interpretation method, after expanding the range of features, a small number of users still chose not to agree. However, most users still disagreed with the explanation given by the Mask explanation method

Figure 15. Sankey Diagram results for Gradient-weighted Class Activation Mapping (Grad-cam). The third column shows the distribution of top-3% of user satisfaction, and the fourth column shows the distribution of top-8% of user satisfaction. Grad-cam method is tested with 10 images from DataS and 35 users.

Figure 16. Sankey Diagram results for Mask. The third column shows the distribution of top-3% of user satisfaction, and the fourth column shows the distribution of top-8% of user satisfaction. Mask method is tested with 10 images from DataS and 35 users.

Figure 17. Sankey Diagram results for Rise. The third column shows the distribution of top-3% of user satisfaction, and the fourth column shows the distribution of top-8% of user satisfaction. Rise method is tested with 10 images from DataS and 35 users.

4.4. Evaluation of Class-Discriminativeness

In this section, we aimed to compare three different visual explanation methods in terms of their power of discrimination across evaluation dataset. Table 4 shows the discrimination of the three different methods in DataT. We can see that the Grad-cam and Rise method had a failure case respectively. For example, in case 5.1, the predicted result of the VGG-16 model was banana, and the visual explanation given by the Grad-cam method focused on the feature of pineapple. In case 4.1, the explanation given by the Rise method focused on both the mouse and keyboard features, when the model predicts that the input was a mouse. Except for the success of the Mask method in case 4.2 (keyboard prediction result), all other tests showed that it was not class discrimination. The corresponding case display above can be seen in Table 5. In comparison, we found that the Mask method failed to have class discrimination under the DataT dataset.

Table 4. Comparisons of class discrimination of three different explanation methods. The first column is the five test examples of the DataT. The second column is the prediction result of the pre-trained VGG-16 model. The last three columns are the class discrimination results of the explanation methods. (✓ mark indicates that the class-discriminateness characteristic is satisfied).

Sample	Case	Prediction Result	Grad-CAM	Mask	Rise
1.1	Pineapple	✓	✗	✓
	1.2	Cat	✓	✗	✓
2.1	Knife	✓	✗	✓
	2.2	Broccoli	✓	✗	✓
3.1	Football	✓	✗	✓
	3.2	Helmet	✓	✗	✓
4.1	Mouse	✓	✗	✗
	4.2	Keyboard	✓	✓	✓
5.1	Banana	✗	✗	✓
	5.2	Pineapple	✓	✗	✓

Table 5. Cases in class-discriminateness of different interpretation algorithms (P means the model prediction result).

Test Sample
	P = Banana	P = Keyboard	P = Mouse
Interpretation Result
	Grad-cam	Mask	Rise

To further examine the class discrimination of these explanation methods, we also compared the class discrimination of these methods under DataS dataset. The results are shown in Table 6. The point game value of the Grad-cam and Rise methods were both 100%, but Mask had eight failed cases in the 40 samples. As shown in the Figure 18, in the failure cases of pandas, goats, and school bus, the explanation given by the Mask method was spread over the entire image pixel, in which the foreground and background were identical importance. For the uniform failure, the Mask method explanation method focused on the soldier's jaw rather than on uniform features.

Figure 18. Failure cases of Mask method in panda category, goat category, school bus category and uniform category. (Note that these examples are from DataS.)

Table 6. Point game statistics of three explanation methods on DataS.

Interpretation Methods	Grad-Cam	Mask	Rise
Number of missed objects	0	8	0
Point Game Value	100%	80%	100%

Table 7 shows the result of point games for these methods under DataM dataset. All methods could successfully hit the object itself in the image. In order to further measure the discriminative performance of different explanation methods in DataM, we counted the number of objects hit by these methods, and the results are reported in Figure 19. We found that Mask method could highlight all objects related to the predicted category and find as much evidence as possible to corroborate the predicted results. At the same time, we investigated user satisfaction with these three interpretation methods. Figure 20 shows the distribution of user satisfaction on explanation persuasibility. We showed each participant in turn the top 8% occlusion map of the explanation method corresponding to all the images in the DataM dataset, which each user only tested one explanation method to ensure the credibility of the experimental results. We can see that the Mask method had the most votes at the 1 level. This indicates that although the mask method found all possible evidence, these explanations results could not convince users. The reason is that this method failed to successfully give the most relevant features of the predicted results.

Table 7. Point game statistics of three explanation methods on DataM.

Interpretation Methods	Grad-cam	Mask	Rise
Number of missed objects	0	0	0
Point Game Value	100%	100%	100%

Figure 19. The statistical result of the number of objects hit by different interpretation algorithms. Each interpretation method is tested with 10 images from DataM and 35 users.

Figure 20. The user satisfaction distribution of different interpretation algorithms. Each interpretation method is tested with 10 images from DataM and 35 users. Satisfaction score is 1–5 (1 means most dissatisfied, 5 means most satisfied).

4.5. Evaluation Summary

Table 8 summarizes how the visual explanation methods perform on different evaluation perspective. We can see that Grad-cam had high correctness property, high persuasibility and high class discriminativeness. Mask had low correctness property, and the persuasibility was also low. To our surprise, this interpretation method did not have the class discriminativeness property. Compared with the other two methods, Rise had medium correctness and persuasibility property. At the same time, it had high class discriminativeness property.

Table 8. Comparison of the considered visual explanation method. Different marks are provided: “H” (High), “M” (Medium), “L” (Low) or NA (Not Applied).

Explanation Methods	Correctness	Persuasibility	Class Discriminativeness
Grad-cam	H	H	H
Mask	L	L	NA
Rise	M	M	H

5. Discussion

The aim of this paper is to evaluate the reliability of visual interpretation methods. A hierarchical evaluation dataset is constructed based on category labels and all data is selected from ImageNet. We can supplement from PASCAL VOC07 [50] to further enrich the evaluation data set. Although we only compared three classic visual interpretation methods, our method has good universality and can be extended to other visual interpretation methods. The disadvantage of the experimental setting is that the occlusion parameter setting of visual features and the selection of the number of volunteers are empirically determined. In the next step, the bias data set can be added to explore whether the visual interpretation method explicitly finds the bias pattern of the model.

We use a Venn diagram (Figure 21) to show the relationship between these three dimensions. D2 (persuasibility) and D1 (accuracy) have overlapping parts, but they are not exactly the same. The persuasive experiment wants to verify whether the user can recognize the object when the relevant feature blocks (calculated by the interpretation algorithm)

appear at the same time. The accuracy experiment further evaluates the interpretation algorithm by allowing users to subjectively select feature blocks to obtain feature importance. D3 (Class-discriminateness) is one of the most basic attributes. The principle of the evaluation algorithm is to hit the object target. The above three evaluation dimensions belong to the category of plausibility, and the ultimate goal is to evaluate which kind of visual interpretation is more convincing.

Figure 21. The Venn diagram of D1 (accuracy), D2 (persuasibility) and D3 (class-discriminateness).

6. Conclusions

In this paper, we propose and demonstrate an evaluation framework of interpretation methods. We focus on the image classification domain and adopt the user subjective evaluation method to investigate the performance of interpretation methods at the user level. Our experiments reveal the usefulness of different interpretation methods from three dimensions, producing a more realistic evaluation of saliency map. Through human studies and explanations, we have proven that the performance of existing visual interpretation methods is overestimated. These results open an exciting perspective on designing visual interpretation principles. Our method is also a starting point, and we hope that our work can contribute to the evaluation of visual interpretation methods.

Author Contributions: Conceptualization, J.L., D.L. and Y.W.; methodology, J.L.; software, J.L.; validation, J.L., D.L. and Y.W.; formal analysis, J.L.; investigation, J.L.; resources, J.L.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L., D.L. and Y.W.; visualization, J.L.; supervision, G.X. and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Aerospace Information Research Institute.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patients to publish this paper.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in <https://github.com/muzi-8/Reliable-Evaluation-of-XAI>, accessed on 10 March 2021.

Acknowledgments: The authors also would like to thank the anonymous reviewers. Last but not least, this paper is not the work of a single person but of a great team including members of Key Laboratory of Network Information System Technology.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Round 1

Question
Can you recognize this object?
Image Display

Answer
☐ Yes ☐ No
(If the user checks the yes button, the system jumps to the following interface and ends, otherwise the system enters round 2)
Please give the name of the object you think

Round 2

Question
This is actually an elephant, do you agree?

Answer
☐ 1-Strongly disagree ☐ 2-Disagree ☐ 3-Generally agree ☐ 4-Agree ☐ 5-Strongly agree
(If the user selects the 1 and 5 button, it will end directly, otherwise the system enters round 3)

Round 3

Question
Look again, do you agree that it is an elephant this time?

Answer
☐ 1-Strongly disagree ☐ 2-disagree ☐ 3-Generally agree ☐ 4-agree ☐ 5-Strongly agree

Figure A1. User satisfaction evaluation interface.

Appendix D

Figure A4. Examples of ground truth result of the relevant feature block marked by the user of each category. (TOP) Ten input image samples in ten categories from the evaluation data set DataS. (Middle) Segmented image with 100 feature blocks obtained according to the SLIC segmentation algorithm. (Bottom) Resulting multi-participants human attention mask. Each image is annotated by 40 unique participants.

Appendix E

Figure A5. Examples of ground truth result of the relevant feature block marked by the user of each category. (TOP) Ten input image samples in ten categories from the evaluation data set DataS. (Middle) Segmented image with 100 feature blocks obtained according to the SLIC segmentation algorithm. (Bottom) Resulting multi-participants human attention mask. Each image is annotated by 40 unique participants.

References

1. Haenlein, M.; Kaplan, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *Calif. Manag. Rev.* **2019**, *61*, 14–5. [[CrossRef](#)]
2. Yu, H.; Siddharth, N.; Barbu, A.; Siskind, J. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *J. Artif. Intell. Res.* **2015**, *52*, 601–713. [[CrossRef](#)]
3. Sorower, M.S.; Slater, M.; Dietterich, T.G. Improving Automated Email Tagging with Implicit Feedback. In Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, Charlotte, NC, USA, 11–15 November 2015; pp. 201–211.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
6. Chiu, C.C.; Sainath, T.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, K.; et al. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AL, Canada, 15–20 April 2018; pp. 4774–4778.

7. Zhu, Q.; Zhou, X.; Wu, J.; Tan, J.; Guo, L. *A Knowledge-Aware Attentional Reasoning Network for Recommendation*; AAAI: Menlo Park, CA, USA, 2020.
8. Yang, Z.; Zhang, A.; Sudjianto, A. Enhancing Explainability of Neural Networks through Architecture Constraints. *IEEE Transact. Neural Netw. Learn. Syst.* **2020**. [\[CrossRef\]](#)
9. Camburu, O.M. Explaining Deep Neural Networks. *arXiv* **2020**, arXiv:2010.01496.
10. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intell.* **2019**, *267*, 1–38. [\[CrossRef\]](#)
11. Liang, H.; Ouyang, Z.; Zeng, Y.; Su, H.; He, Z.; Xia, S.; Zhu, J.; Zhang, B. *Training Interpretable Convolutional Neural Networks by Differentiating Class-specific Filters*; European Conference on Computer Vision. Springer, Cham, 2020: 622–638.
12. Zhang, Q.; Wu, Y.; Zhu, S. Interpretable Convolutional Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8827–8836.
13. Wagner, J.; Köhler, J.; Gindele, T.; Hetzel, L.; Wiedemer, J.T.; Behnke, S. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9089–9099.
14. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.
15. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–18 August 2016.
16. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [\[CrossRef\]](#)
17. Fong, R.C.; Vedaldi, A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3449–3457.
18. Petsiuk, V.; Das, A.; Saenko, K. *RISE: Randomized Input Sampling for Explanation of Black-Box Models*; BMVC: Cardiff, UK, 2018.
19. Ghorbani, A.; Abid, A.; Zou, J.Y. *Interpretation of Neural Networks is Fragile*; AAAI: Menlo Park, CA, USA, 2019.
20. Subramanya, A.; Pillai, V.; Pirsiavash, H. Fooling Network Interpretation in Image Classification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 22 April 2019; pp. 2020–2029.
21. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
22. Antunes, P.; Herskovic, V.; Ochoa, S.; Pino, J. Structuring dimensions for collaborative systems evaluation. *ACM Comput. Surv.* **2008**, *44*, 8:1–8:28. [\[CrossRef\]](#)
23. Freitas, A. Comprehensible classification models: A position paper. *SIGKDD Explor.* **2014**, *15*, 1–10. [\[CrossRef\]](#)
24. Kim, B.; Chacha, C.M.; Shah, J. Inferring Robot Task Plans from Human Team Meetings: A Generative Modeling Approach with Logic-Based Prior. *arXiv* **2013**, arXiv:1306.0963.
25. Khakzar, A.; Baselizadeh, S.; Khanduja, S.; Kim, S.T.; Navab, N. Explaining Neural Networks via Perturbing Important Learned Features. *arXiv* **2019**, arXiv:1911.11081.
26. Yang, M.; Kim, B. Benchmarking Attribution Methods with Relative Feature Importance. *arXiv* **2019**, arXiv:1907.09701.
27. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2013**, arXiv:1312.6034.
28. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M.A. Striving for Simplicity: The All Convolutional Net. *arXiv* **2014**, arXiv:1412.6806.
29. Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
30. Qi, Z.; Khorram, S.; Li, F. Visualizing Deep Networks by Optimizing with Integrated Gradients. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019.
31. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014.
32. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [\[CrossRef\]](#)
33. Guidotti, R.; Monreale, A.; Turini, F.; Pedreschi, D.; Giannotti, F. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv. (CSUR)* **2019**, *51*, 1–42. [\[CrossRef\]](#)
34. Liang, Y.; Li, S.; Yan, C.; Li, M.; Jiang, C. Explaining The Black-Box Model: A Survey of Local Interpretation Methods for Deep Neural Networks. *Neurocomputing* **2021**, *419*, 168–182. [\[CrossRef\]](#)
35. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Interpretable machine learning: Definitions, methods, and applications. *arXiv* **2019**, arXiv:1901.04592.
36. Samek, W.; Wiegand, T.; Müller, K. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv* **2017**, arXiv:1708.08296.
37. Zhang, Q.; Zhu, S. Visual interpretability for deep learning: A survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 27–39. [\[CrossRef\]](#)
38. Jeyakumar, J.; Noor, J.; Cheng, Y.H.; Garcia, L.; Srivastava, M. *How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods*; NeurIPS: San Diego, CA, USA, 2020.

39. Hooker, S.; Erhan, D.; Kindermans, P.; Kim, B. Evaluating Feature Importance Estimates. *arXiv* **2018**, arXiv:1806.10758.
40. Oramas, J.; Wang, K.; Tuytelaars, T. Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks. *arXiv* **2019**, arXiv:1712.06302.
41. Zhou, B.; Bau, D.; Oliva, A.; Torralba, A. Interpreting Deep Visual Representations via Network Dissection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2131–2145. [[CrossRef](#)]
42. Linsley, D.; Eberhardt, S.; Sharma, T.; Gupta, P.; Serre, T. What are the Visual Features Underlying Human Versus Machine Vision? In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 29 October 2017; pp. 2706–2714.
43. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. *ImageNet: A Large-Scale Hierarchical Image Database*; CVPR: Salt Lake City, UT, USA, 2009.
44. Ghorbani, A.; Wexler, J.; Zou, J.; Kim, B. *Towards Automatic Concept-Based Explanations*; NeurIPS: San Diego, CA, USA, 2019.
45. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [[CrossRef](#)]
46. Krzywinski, M.; Altman, N. Visualizing samples with box plots. *Nat. Method.* **2014**, *11*, 119–120. [[CrossRef](#)] [[PubMed](#)]
47. Soundararajan, K.; Ho, H.K.; Su, B. Sankey diagram framework for energy and exergy flows. *Appl. Energy* **2014**, *136*, 1035–1042. [[CrossRef](#)]
48. Joshi, A.; Kale, S.; Chandel, S.; Pal, D. Likert Scale: Explored and Explained. *Br. J. Appl. Sci. Technol.* **2015**, *7*, 396–403. [[CrossRef](#)]
49. Zhang, J.; Lin, Z.; Brandt, J.; Shen, X.; Sclaroff, S. Top-Down Neural Attention by Excitation Backprop. *Int. J. Comput. Vis.* **2017**, *126*, 1084–1102. [[CrossRef](#)]
50. Everingham, M.; Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2009**, *88*, 303–338. [[CrossRef](#)]