The Visual Computer (2023) 39:6097–6113
https://doi.org/10.1007/s00371-022-02715-8

ORIGINAL ARTICLE

Visual explanation and robustness assessment optimization of saliency maps for image classification

Xiaoshun Xu[1,2] · Jinqiu Mo[1]

Accepted: 20 October 2022 / Published online: 12 November 2022

Abstract
For image classification using Deep Learning, applying visual explanations allows end-users to understand better the basis of model decisions in the inference process. Our method optimizes the black-box visual explanation called Randomized Input Sampling for Explanation (RISE) by proposing the concept of Decisive Saliency Map (DSM) and the corresponding quantitative metric. The introduction of DSM makes the discriminative salient regions more prominent and easier to understand with ignorable extra costs. Moreover, DSM efficiently correlates robustness assessment with the visual explanation via saliency value distribution. It provides a reference indicator for the reliability and robustness assessment of the model predictions, complementing the common-used Softmax confidence score. Experiments demonstrate that the utilization of DSM and the related quantitative metric can improve the visualization of mainstream CNN models, and differentiate the concrete importance of confusingly similar salient regions. By quantitatively assessing the robustness of the inference process, DSM identifies the potential misclassification risk of high-performance CNN models accurately.

Keywords Visual explanation · Data analysis · Robustness assessment · Black-box models

1 Introduction

As an essential branch of Machine Learning proliferating, Deep Learning can achieve high-performance, multi-purpose machine vision applications by skillfully designing convolutional neural networks (CNN) models and training with adequate image samples. Nowadays, CNN applications are gaining popularity for quality control of products in automatic manufacturing environments. Especially image classification tasks realize the most prevalent CNN-based error-proofing machine vision applications [1, 2]. However, as a subset of Machine Learning, CNN has the same shortcomings. It is difficult to understand the decision mechanism and is not as intuitive as pattern-based machine vision solutions with hand-crafted rules. Usually, end-users tend to be more skeptical of CNN-based applications than pattern-based machine vision solutions [3].

With the superior performance of Machine Learning and the increasing number of applications in many fields, the need to improve its interpretability has become mandatory. Explainable Artificial Intelligence (XAI) is the concept that Machine Learning models are required to be interpretable, trustworthy, and efficiently manageable [4, 5]. Since its emergence, XAI has been widely emphasized by the government, academia, and industry. Through the joint efforts of professional organizations, XAI-related ISO standards [6] have been gradually established and published.

To provide end-users with a better understanding and trust of image classification tasks based on CNN, the rational implementation of visual explanations for inference mechanisms is a feasible and reliable solution. When end-users can understand a black-box model's decision process to assess or verify the output via visualization, they will be receptive to model deployment [3, 7–9]. Proper visual explanations also localize fine-grained features of the image [10], estimate the influences of the wrong prediction, reveal deficiencies in the input data and training process, and offer guidance for continuous updates and efficient improvement of the models [11]. In real-world medical applications, IGOS + + [12] discovered the classification model overfitting to the texts instead of the indicative symptoms of pneumonia on the X-ray images. For

✉ Xiaoshun Xu
xxs_sgm@aliyun.com

[1] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[2] SAIC General Motors Corporation Limited, Shanghai 201206, China

CNN applications in industry, visual explanations facilitate avoiding quality control incidents in massive manufacturing due to the lack of model transparency and interpretability [7].

Another essential property of interpretability for machine vision in industrial quality control applications is robustness assessment. High robustness means that the model can maintain its regular performance under perturbation or worse situations [6, 13], i.e., minor changes in the input should not cause significant variance in the model output. While reflecting and understanding the interpretability of model inferences, assessing and verifying the robustness of model applications has also become a well-defined workflow required by ISO standards. Typical industrial manufacturing cases are deployed in physically enclosed environments and secure local networks. Although malicious intrusion in adversarial attacks is almost impossible, machine vision applications are often subject to unexpected disturbances from external working conditions. Typical disturbances are occlusions in ROI (region of interest), greyscale brightening (or darkening), blurring, and other feature contaminations caused by environmental degradation, workpiece, manufacturing process, and equipment. Robust machine vision applications should rigorously ensure that the inference results do not lead to false positives when affected by small amounts of typical disturbances. Also, the quantity of false negatives is within an acceptable range.

Early visual explanation research focused on analyzing the model structure and visualizing the features processed in various network layers [14–17]. These methods required the models to be white-box. Some visualizations were not easily understandable to end-users but merely interpretable for researchers. With the promotion of CNN technology and the support of XAI program, plenty of visual explanations [18–21] applicable to black-box models have been proposed and proven effective. Still, they do not incorporate the function of robustness assessment and validation, also some minor insufficiency to be improved for practical applications.

Regarding academic advances and industrial needs, our paper optimizes RISE [20] for better applicability to CNN-based image classification tasks by proposing the concept of Decisive Saliency Map and the corresponding quantitative metric. Our method derives appropriate threshold values and weights based on the characteristics of saliency value distribution, then performs binarization and weighted sum-up operations for the feature regions with the highest importance to obtain DSM and its coverage rate. The main contributions of this paper are as follows:

We carried out the data analysis of the distribution of the saliency value to propose the concept of DSM. The characteristic of the distribution is merged into the visual explanation to highlight the essential salient regions that determine model inference. Several comparisons display the differences in feature graininess focused by CNN models of different depths by applying DSM.

Our optimization method continues the concise and easy-to-implement ideas of RISE, making the visual explanation more intuitive but less dispersive. Our method displays more fine-grained and decisive salient regions for image classification applications via visualization.

The coverage rate of DSM can provide the quantitative robustness assessment and an extra reference indicator of the trustworthiness of the model predictions, complementing the Softmax confidence score. The rate detects samples with high confidence scores but implies a high risk of model misclassification. Furthermore, the robustness assessment we proposed can be adopted into black-box models in industrial machine visions at an ignorable computational cost, responding to the incoming ISO standard requirement for the AI industry.

2 Related work

Due to the differences in research perspectives and purposes, dozens of visual explanations have been developed recently for CNN models with distinct techniques and visual effects [22]. Commonly used Deep Learning visual explanations are generally subordinate to Local Interpretability of Post-Hoc Explainability strategies. They are classified as model-specific and model-agnostic to distinguish visual explanation types. Model-specific methods are designed for specified models of which the designer has a certain level of knowledge at least. Model-agnostic methods are intended for any unknown models or algorithms. In most model-agnostic cases, only these models' inputs (samples) and outputs (predictions) are visible and accessible.

2.1 Model-specific methods

Model-specific methods are the most interpretability techniques contributed by the AI community, and always provide reliable and fundamental explanations. Model-specific visual explanations for CNN are based on backpropagation and class activation mapping (CAM), including Guided Backprop [14], CAM [23], GRAD-CAM [24], Score-CAM [25] etc. Although these methods have excellent and understandable visualizations, they require accessing or modifying partial model network layers to perform specific operations, including global averaging and weighted summation of gradients, class activation maps, weights of convolutional feature maps, or forward passing scores on object classes [26, 27].

Therefore, model-specific approaches do not apply to complex CNN models deployed after encapsulation, e.g., mainstream industrial machine vision products.

2.2 Model-agnostic methods

Typical model-agnostic methods are based on local approximate interpretability, or sensitivity analysis of how the output is influenced by perturbed input [11, 28].

Local approximate interpretability constructs simplified models to explain linearly, such as Local Interpretable Model-agnostic Explanations (LIME) [18] and its variant Anchor [21].

Though both are based on superpixels segmentation, Anchor improves LIME by anchoring and superimposing with if–then rules. However, the prerequisite for establishing an Anchor interpretation of image classification is to obtain the correct superpixels segmentation, which may lead to considerable variances in visualization due to particular segmentation algorithms and hyper-parameters. It also requires that the image sample has adequate discriminative feature areas to build a reasonable explanation for Anchor.

The sensitivity analysis method generates the saliency maps of the input image by analyzing changes in the prediction influenced by the input perturbation. Typical saliency maps used for CNN models [27] are heat maps masked on the original input images, reflecting the different degrees of influence of the feature regions by heat colours. Representative methods as [19] and [20]. They visualize the image feature regions that significantly influence the results and include quantitative metrics of accuracy on interpretability: deletion and insertion. RISE [20] already possesses concise design, excellent performance, and high applicability. These properties are the preliminary basis for the feasibility of applications in industrial environments. Recently, based on the overall experimental results of several benchmark datasets and CNN models, RISE is one of the best methods evaluated by five recognized visualization metrics [29].

However, there is still space for improvement in reflecting visualization attributes, particularly feature graininess and importance ranking [30, 31]. Meaningful perturbation [19] requires additional meta-parameters but provides less sharp visualization due to Gaussian blur masking. It is difficult to identify the feature importance since it only contains coarse feature patterns. RISE may confuse judgments due to minor salient regions and noises generated by the randomness of the masking process. By balancing the advantages of deletion and preservation processes, the recent study IGOS + + [12] uses bilateral perturbations to generate fine-grained saliency maps with additional cost. Meanwhile, its scattered salient regions may interfere with subjective cognition.

Although the available visual explanations help understand and highlight the feature regions, not many are applicable to black-box models while having explicit effects. To our knowledge, none of them assess the robustness of the model inference process in conjunction with the visualizations.

3 Decisive saliency map (DSM)

Motivated by RISE, after measuring the differences in predictions by using thousands of perturbed input images and obtaining the saliency maps for each label class, we further utilize the saliency value distribution information to improve the visual explanation. DSM is calculated to highlight the essential feature regions. It suppresses the dispersion [29], noises, and minor feature distractions from the random masking process. The coverage rate of DSM serves as a quantitative metric and comparison criteria.

The overall flowchart is shown in Fig. 1.

3.1 Acquisition of decisive saliency maps

In CNN models, for a 3-channel input image $I \in R^{H\times W\times 3}$ with length H and width W,$f(I)$ is the confidence score (probability) of the inference on the input and processed by the Softmax function. As defined in (1) [20], $f(I\odot M)$ is the confidence score of the perturbed images obtained after the element-wise random multiplication operation of the original image. $M\in\{0,\ 1\}$ is the random binary mask, with the probability p for unmasking. It is empirically set to 0.5 from the range [0,1], indicating half image patches occlusion.$\mathbb{E}[M]$ is the expectation value of all possible masking operations when the image pixel λ is still preserved as $M(\lambda)=1$.MC denotes a total of N times of Monte Carlo sampling masking operations and model inferences. The importance of each pixel is approximately computed by the weighted average of the masks and corresponding confidence scores, then generates saliency maps $S_{I,f}(\lambda)$ to the inferences as follows:

$$S_{I,f}(\lambda)\overset{\mathrm{MC}}{\approx}\frac{1}{\mathbb{E}[M]\cdot N}\sum\nolimits_{i=1}^{N}f(I\odot M_i)\cdot M_i(\lambda) \tag{1}$$

For a CNN model with K classes, the stacked 1-channel saliency maps obtained from (1) is $S_{I,f}(\lambda)\in R^{H\times W\times 1\times K}$, represented as S^K. For a specific class k in the saliency map $S^k\in R^{H\times W\times 1}$, the values of the saliency map for each pixel are scalars $s_{ij}\in(0,\ 1)$ indexed by i, j in height and width, respectively.

Let the maximum, minimum, and mean values of s_{ij} in S^k be $max(s_{ij})$,$min(s_{ij})$, and $mean\big(s_{ij}\big)$, respectively. According to (1), it can be derived intuitively that s_{ij} has the following properties:

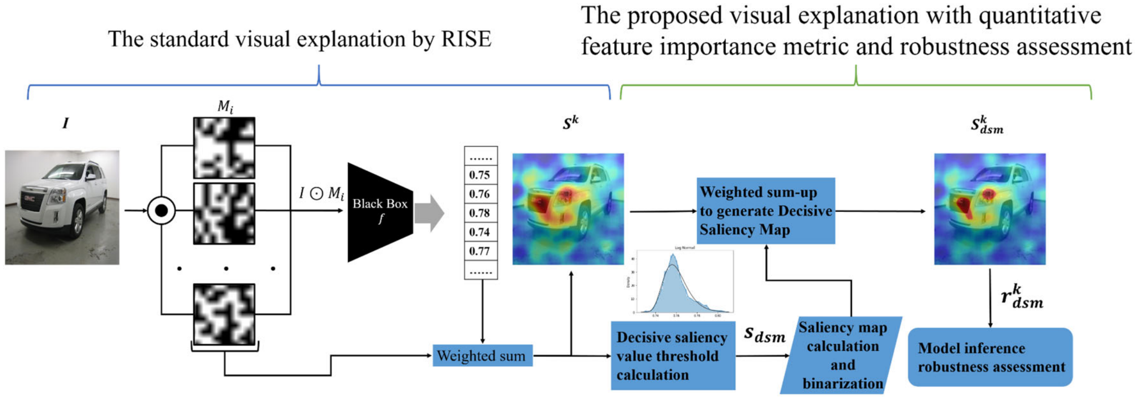

The standard visual explanation by RISE
The proposed visual explanation with quantitative feature importance metric and robustness assessment
I
M_i
$I \odot M_i$
Black Box f
......
0.75
0.76
0.78
0.74
0.77
......
S^k
Weighted sum
Weighted sum-up to generate Decisive Saliency Map
S^k_{dsm}
Decisive saliency value threshold calculation
S_{dsm}
Saliency map calculation and binarization
r^k_{dsm}
Model inference robustness assessment

Fig.1 Overall flowchart of the proposed method

1. The value of s_{ij} is positively correlated with the classification confidence score $f_k(I)$ of the original image, and the number of high-importance pixels preserved by the random masking operations $\mathrm{M_i}(\lambda)$. The confidence score $f_k(I \odot M_i)$ of the masked image is smaller than the original image confidence score $f_k(I)$ in most cases.
2. Since the masked input and the model for the confidence score $f_k(I \odot M_i)$ are identical in the saliency map of the same class, the absolute value difference of $\max(s_{ij})$ and $\min(s_{ij})$ mainly comes from the total number of saliency pixels preserved by the mask $M_i(\lambda)$ that can reproduce most class feature information.
3. The larger the values of $\max(s_{ij})$, $\min(s_{ij})$, $\mathrm{mean}\left(s_{ij}\right)$ and the closer to $f_k(I)$, the higher the number of pixels in the high-saliency regions. It means the inferences can hardly be perturbed into misclassification by the random masks. Accordingly if $f_k(I \odot M_i)$ is generally higher, the robustness of the inference process is relatively better in industrial applications. And vice versa, low $f_k(I \odot M_i)$ means the model is more susceptible to random masks with poor robustness. Masking a small part of the salient region leads to a significant decrease in $\max(s_{ij})$ and $\min(s_{ij})$.

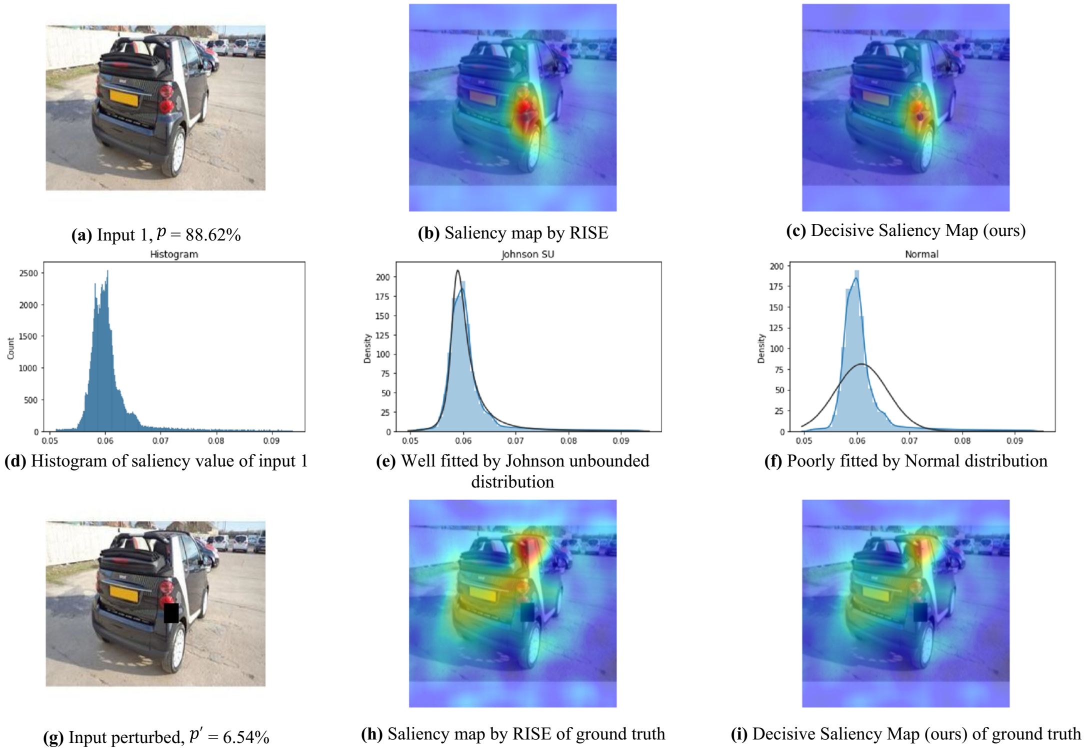
The s_{ij} values of pixels in the saliency map were converted into a histogram for subsequent analysis. Most of the input images with correct inference and excellent confidence scores $f(I)$ have $\max(s_{ij})$ values close to $f(I)$. The distribution histograms Skewness and Kurtosis of saliency maps are relatively small.

However, the analysis also reveals some abnormal input samples. Their model inferences and corresponding saliency maps by RISE are both correct, as in Fig. 2a–b. Though the confidence scores are not low in value probably, their $\max(s_{ij})$ is much smaller than $f(I)$. Figure 2 d–f shows the saliency value distribution fitted by functions. The histogram density distribution with much larger Skewness and Kurtosis values is almost impossible to be fitted by a standard normal distribution. It can only be well fitted using the Johnson unbounded distribution [32]. Such histogram distribution tends to have a significant long-tail effect, and the mean saliency value is too small or even approximate to zero.

Furthermore, the salient regions in the abnormal sample are too small for the inference to be robust. Once a slight perturbation degrades the input image, for instance, brightening or darkening, or physically the object is occluded or rotated [6], as in Fig. 2g–h, the prediction could be misclassified, or the confidence score could be significantly reduced. Even the inference processes of these samples by specific models are risky if they occur in real-world applications, it is difficult for them to be noticed, understood, and accepted by end-users, e.g., industrial quality control personnel.

According to the analysis mentioned above, if the statistical information of s_{ij} data characteristics of the saliency map S^k can be utilized and reflected in the visualization of the saliency map, it can facilitate direct observation and reduce extra computational data records during application. Besides that, it is more feasible to assess the robustness of the model inference based on the input perturbation methods other than the backpropagation approaches from the principle. Therefore, motivated by RISE, we propose an optimized method for the saliency map. In this method, the data characteristics of the saliency maps are merged into the visual explanation through algorithmic transformation. We define the new saliency map as Decisive Saliency Map, which indicates that the feature area covered by the transformed salient region has the dominant influence and decisive effect on the image classification prediction. Weighting decisive salient regions into the heat map can correctly correlate with the data characteristics of the saliency distribution histogram to improve the visualization of the importance of features and provide a reliable robustness assessment.

The process of computing the decisive saliency maps S^K_{DSM} is as follows:



(a) Input 1, $p = 88.62\%$

(b) Saliency map by RISE

(c) Decisive Saliency Map (ours)

(d) Histogram of saliency value of input 1

(e) Well fitted by Johnson unbounded distribution

(f) Poorly fitted by Normal distribution

(g) Input perturbed, $p' = 6.54\%$

(h) Saliency map by RISE of ground truth

(i) Decisive Saliency Map (ours) of ground truth

Fig. 2 Visualizations of saliency maps of a typical example of poor robustness for model inference. The input sample **a** is from Stanford Car-196 [33] test set, and the CNN model is a fine-tuned EfficientNet-B3 [34] with transfer learning. The saliency map provided by RISE in **b** does not reflect the saliency value distribution characteristics. However, the saliency value distribution histogram has an apparent long-tail effect in **d–f**. Once minor disturbances in **g** perturb the input, the classification result is incorrect, or the score is unexpectedly low. Our DSM method can reflect the risk of insufficient robustness of model inference by visualizing the decisive salient region in **c** and **i**

Step 1 Obtain the stacked saliency maps $S^K \in R^{H \times W \times 1 \times K}$ of the CNN model for image classification with K classes using (1).

Step 2 Select the two-dimensional saliency map $S^k \in R^{H \times W \times 1}$ for the k^{th} class as needed, search for its $\max\left(s_{ij}\right)$ and $\min(s_{ij})$, calculate the mean value $\text{mean}\left(s_{ij}\right)$. The decisive saliency differential value δ_s is derived to reflect the severity of the long-tail effect of the histogram distribution using the data characteristics of the saliency map according to (2):

$$\delta_s = C_d\left(\max\left(s_{ij}\right) - \min\left(s_{ij}\right)\right)\frac{\text{mean}\left(s_{ij}\right)}{\min\left(s_{ij}\right)} \tag{2}$$

where C_d is set as the coefficient of dominance. It is used to appropriately distinguish the importance of the original image features and effectively suppress the noises in the subsequent heat map without completely ignoring the subordinated features.

The range of C_d value practically meaningful for the subsequent binarization operation is $C_d \geq 0$. In [35], the research suggests a normalized weight threshold to select a highlighted region for occlusion to improve robustness during training. In our design concept, C_d should not only separate the decisive salient regions from the rest of the image but also properly suppress subordinate features.

The value of C_d in range of [0.1, 0.5] is proposed initially regarding the simplicity of the visualization design and interpretation function [36]. The improper value of C_d would make it difficult to distinguish between differences in the robustness of the inference, or excessively ignore subordinate visualized features. When the value of C_d is set to 0.2, the following equations can equivalently emphasize the decisive salient regions with explicit boundaries and straightforwardly skip the normalization process of saliency values.

Step 3 Calculate the saliency threshold s_{dsm} for the subsequent binarization operation on s_{ij}. Compared with the linear or fixed coefficient operation of $\max\left(s_{ij}\right)$ as the binarization threshold, the use of exponential form $e^{-\delta_s}$ can accurately distinguish the influence level of pixels in the saliency map and the distribution characteristics of the saliency value histogram:

$$s_{dsm} = \max(s_{ij}) \cdot e^{-C_d(\max(s_{ij})-\min(s_{ij}))\cdot\frac{\text{mean}(s_{ij})}{\min(s_{ij})}} \tag{3}$$

Step 4 Binarize all s_{ij} in S^k with s_{dsm} as the threshold to obtain a new saliency map $\widetilde{S}^k \in R^{H\times W\times 1}$ consisting of $\widetilde{s}_{ij}$:

$$\widetilde{s}_{ij} = \begin{cases} 1, \, if \quad s_{ij} > s_{dsm} \\ 0, \, if \quad s_{ij} \leq s_{dsm} \end{cases} \tag{4}$$

Step 5 Sum up of S^k and the weighted $\widetilde{S}^k$. The weight is δ_s corresponding to the k^{th} class. The selected regions are emphasized with the contribution of C_d. Our converged visualization meets the focal point principle related to human attention well [36]. This process is equivalent to superimposing a small portion of the saliency value of essential features on the original heat map. The optimized visualization meets the closure principle that patterns should be clustered with definite borders when visual explanation contains complex feature elements. Thus, Decisive Saliency Map S^k_{dsm} for single class and $S^K_{DSM} \in R^{H\times W\times 1\times K}$ for all classes are obtained as follows:

$$S^k_{dsm} = S^k + \delta_s \cdot \widetilde{S}^k \tag{5}$$

$$S^K_{DSM} = \{S^1_{dsm}, S^2_{dsm}, S^3_{dsm}, \ldots\ldots, S^k_{dsm}\} \tag{6}$$

Our method merges the saliency maps with implicit data characteristics information of saliency value histograms, and represents the fine-grained features by delineating the realistic feature boundary. The optimization is still simple and effective in design concepts. For the original and perturbed input sample 1, Decisive Saliency Maps are shown in Fig. 2c and Fig. 2i. Since the highlighted area provided by DSM is almost impossible to be seen in Fig. 2i after perturbation in Fig. 2g, it is more intuitive to explain the unexpectedly low confidence score due to the lack of features than in Fig. 2h.

3.2 DSM-based evaluation metric

Causal metrics have been commonly used in previous research to objectively evaluate the performance of visual explanations, e.g., AUC scores (Area Under probability Curve) of the deletion and insertion, pointing game, etc. These approaches mainly concentrate on validating the accuracy, localization, and faithfulness of the saliency maps. They do not involve the robustness assessment of the model inferences. Also, AUC calculations require additional GPU inferences, increasing computational cost and time extensively.

In image classification tasks based on CNN models, even if the subjective observations of saliency maps of various input samples are similar and the objective AUC calculations are approximate in value, the inference processes still have significant differences regarding the dependence of features, which can be reflected by S^K_{DSM} .

To better analyze the differences in visual explanations, a new quantitative evaluation metric r^k_{dsm} is proposed in this paper, namely the calculation of the coverage rate of DSM. As a concise and intuitive quantitative metric, r^k_{dsm} directly reflects the ratio of pixels of the decisive salient region in an image for the k^{th} class, which directly quantifies the robustness of image classification of black-box models to potential perturbation. Using $\widetilde{s}_{ij}$ from (4) to derive the r^k_{dsm} of specified class from S^k_{dsm} as:

$$r^k_{dsm} = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}\widetilde{s}_{ij} \tag{7}$$

The metric r^k_{dsm} does not rely on subjective cognition while reflecting the quantitive difference in the histogram density distribution of saliency value of similar visual explanations. It can be used for long-term tracking to compare whether the inference processes are robust and unnecessary for updates. The computational runtime is much faster and more efficient than causal metrics that rely on GPUs.

3.3 DSM for robustness assessment

The most common metric for judging the trustworthiness of image classification results is the Softmax confidence score. However, it has been proven [37] that the Softmax confidence score tends to lose calibration as the model structure becomes deeper and more complex, making the model overconfident in the prediction. Even high confidence scores do not ensure the reliability and robustness of the inference process nor truly reflect the likelihood of the correct result.

In [37] also verified that temperature scaling is the simplest and most effective solution for confidence probability calibration without affecting the model's accuracy. The Softmax function σ_{SM}, which converts the network logit vectors $\mathbf{z}_i$ to confidence score at the end of the model networks, is calibrated by adding the temperature parameter T in (8), then the prediction $\widehat{q}_i$ is calibrated as in (9). However, this calibration solution requires access to the model design, making it impossible to apply to most CNN models encapsulated and deployed in the industrial environment.

$$\sigma_{\text{SM}}(\mathbf{z}_i/T)^{(k)} = \frac{\exp\left(z_i^{(k)}/T\right)}{\sum_{j=1}^{K}\exp\left(z_i^{(j)}/T\right)} \tag{8}$$

$$\widehat{q}_i = \underset{k}{max}\,\sigma_{\text{SM}}(\mathbf{z}_i/T)^{(k)} \tag{9}$$

Table 1 Saliency value distribution data of input samples.r^k_{dsm} can be a reference indicator for reliability complementing the confidence score, as most wrong predictions relate to low r^k_{dsm} values but acceptable scores

Input sample number	Dataset/ Source	CNN Model	Name of Class *k*	Prediction Class No.	Probability	GT & Top 1	r_{dsm} for class *k*	Mean saliency	Max saliency	Min saliency	Skewness	Kurtosis
1	Stanford Car-196	EfficientNet-B3	Smart fortwo 2012	196	**88.62%**	**Y**	**0.088%**	0.0597	0.0915	0.0516	3.2850	13.5925
2	ImageNet	ResNet50	goldfish	1	**32.36%**	**Y**	**0.317%**	0.2529	0.2904	0.2248	0.6174	0.2646
		EfficientNet-B0	goldfish	1	**30.90%**	**Y**	**0.016%**	0.0059	0.0073	0.0052	1.3897	4.1776
3	ImageNet	ResNet50	monarch butterfly	323	**96.21%**	**Y**	**2.007%**	0.5636	0.6150	0.5334	0.9057	-0.1665
		EfficientNet-B0	monarch butterfly	323	**43.22%**	**Y**	**0.357%**	0.1887	0.2117	0.1736	0.7170	2.0763
4	Grad-CAM	ResNet50	bull mastiff	243	**58.59%**	**N**	**0.478%**	0.2566	0.3571	0.2245	1.8389	3.4892
			tiger cat	282	**33.73%**		**0.260%**	0.0777	0.1090	0.0633	1.2270	1.7699
		EfficientNet-B0	bull mastiff	243	**7.06%**	**N**	**0.026%**	0.0122	0.0183	0.0104	1.7780	3.0368
			tiger cat	282	**13.23%**		**0.098%**	0.0377	0.0553	0.0298	1.7780	3.6347
5	ImageNet	ResNet50	tusker	101	**57.90%**	**Y**	**0.074%**	0.0324	0.0460	0.0268	1.6356	3.0002
		EfficientNet-B0	tusker	101	**53.90%**	**Y**	**1.194%**	0.2367	0.2636	0.2184	0.7631	0.0784
6	ImageNet	ResNet50	container ship	510	**99.88%**	**Y**	**0.901%**	0.4712	0.5255	0.4325	0.4518	-0.1325
		EfficientNet-B0	container ship	510	**94.92%**	**Y**	**2.483%**	0.6980	0.7436	0.6681	0.7652	0.8604
7	ImageNet	ResNet50	bubble	971	**32.13%**	**Y**	**11.530%**	0.2724	0.3093	0.1607	-1.7625	6.0376
		EfficientNet-B0	bubble	971	**8.34%**	**Y**	**1.800%**	0.2728	0.3046	0.1896	-1.8886	5.1544
8	ImageNet	ResNet50	malamute	249	**67.67%**	**Y**	**0.706%**	0.1918	0.2598	0.1738	2.6362	7.4817
			husky	250	**28.43%**		**0.383%**	0.2115	0.2684	0.1969	2.0962	3.8617
		EfficientNet-B0	malamute	249	**28.22%**	**N**	**0.155%**	0.0665	0.0919	0.0580	2.2263	6.9589
			husky	250	**50.62%**		**0.337%**	0.1949	0.2677	0.1811	2.7755	8.8442
9	ImageNet	ResNet50	valley	979	**1.76%**	**N**	**0.002%**	0.00004	0.00005	0.00004	0.1960	-1.1639
			cliff dwelling	500	**18.93%**		**0.074%**	0.0510	0.0616	0.0454	0.9236	1.1925
			cliff	972	**61.59%**		**0.004%**	0.00021	0.00025	0.00018	0.3341	-0.5846
		EfficientNet-B0	valley	979	**2.61%**	**N**	**0.002%**	0.0019	0.0021	0.0018	0.1629	-0.7993
			cliff dwelling	500	**3.80%**		**0.004%**	0.0013	0.0014	0.0012	0.1240	-0.5712
			cliff	972	**66.68%**		**0.028%**	0.0072	0.0078	0.0066	0.0857	-1.0954
10	Stanford Car-196	EfficientNet-B3	Cadillac SRX 2012	52	**89.62%**	**Y**	**0.870%**	0.6604	0.7841	0.6343	3.1303	12.0562
11	Stanford Car-196	EfficientNet-B3	GMC Terrain 2012	118	**95.27%**	**Y**	**2.570%**	0.7604	0.8069	0.7298	0.8728	0.8938
12	Stanford Car-196	EfficientNet-B3	Ferrari FF Coupe 2012	101	**59.72%**	**Y**	**0.094%**	0.1065	0.1624	0.0983	3.6340	17.3342
13	Stanford Car-196	EfficientNet-B3	Smart fortwo 2012	196	**63.62%**	**Y**	**0.263%**	0.1591	0.2214	0.1459	3.4798	14.3630
14	Stanford Car-196	EfficientNet-B3	Smart fortwo 2012	196	**19.51%**	**Y**	**0.157%**	0.1470	0.2027	0.1305	2.7582	11.7855

We propose that r^k_{dsm} performs as an additional reference indicator for the robustness assessment of CNN models. It can conveniently and intuitively discover the input samples and classes with poor robustness of model inference while avoiding modifying the model structure to calibrate. S^K_{DSM} and r^k_{dsm} of Decisive Saliency Maps reveal potential risks beneath the subjective observation of the saliency maps or AUC calculation, e.g., the improper essential salient region in S^k_{dsm}, low value of r^k_{dsm} corresponding to the prediction class, or unreliable features displayed in fine-graininess.

Such anomalies indicate that the model failed to avoid overfitting during training and is not capable or driven to explore adequate discriminative features. Overfitting leads the model highly susceptible to unpredictable misclassification and unreasonable confidence score fluctuations due to image perturbation and image quality degradation in real-world applications, significantly deteriorating the robustness of the model.

A typical and necessary method for studying the robustness of various vision architectures is the occlusion in salient regions [15], 19. In image classification models, the prerequisite for high prediction scores $f(I \odot M_i)$ of random masked inputs is robust against severe occlusion. A model with excellent robustness indicates that masked inputs are closer to the original input.

$$f(I \odot M_i)_{argmax} \approx f(I) \qquad (10)$$

In a robust inference process, the larger the Softmax confidence score of the $f_k(I)$ inference result, the saliency values $S_{I,f}(\lambda)$ and its related statistical description, i.e., $\max\left(s_{ij}\right)$, will be larger in value consequently. Given the masking probability in (1), $\mathrm{mean}\left(s_{ij}\right)$ positively correlates with the most likely prediction scores of perturbed inputs. Large saliency values and normal distribution will lead to a small binarization threshold after the transformation using (3) and (4). The obtained threshold further allows Decisive Saliency Maps s^k_{dsm} to include more pixels. Then the larger the proportion of the saliency map r^k_{dsm}, as $r^k_{dsm} \propto f_k(I)$ usually. Through distinct perspectives, our method shares logical similarities with the information loss process in [38].

To improve the robustness, the S^K_{dsm} and r_{dsm} of decisive saliency map can serve as an alternative function to confidence probability calibration, which guides the improvement of the model's training dataset or procedure. Typical actions are using Random Erasing [39] or CutMix [40] for data augmentation, introducing the label smoothing function and applying other regularization, etc.

4 Experiments

4.1 Datasets and implementation

The datasets for validating DSM in this paper are ImageNet [41] and Stanford Car-196 [33]. Three types of CNN models established are listed below:

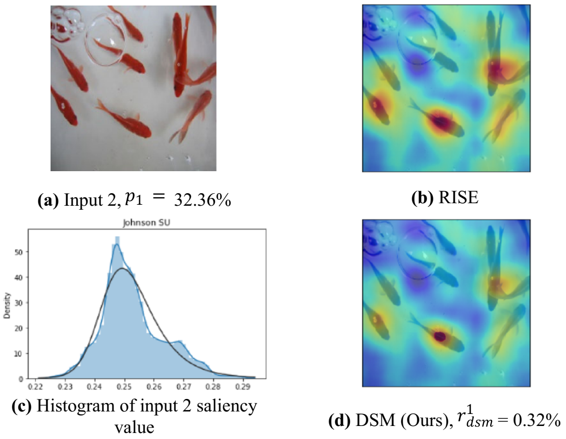
1. **ResNet50** [42], provided by TensorFlow2.3, and its weight pre-trained on the ImageNet dataset. Hereafter referred to as ResNet50.
2. **EfficientNet-B0** [34], provided by TensorFlow2.3, and its weight pre-trained on the ImageNet dataset. Hereafter referred to as EfficientNet-B0.
3. **EfficientNet-B3** [34], which simulates a fine-grained visual classification application deployed in industrial environments, is obtained by transfer learning in TensorFlow from the pre-trained weight on ImageNet with multiple data augmentation, label smoothing, stochastic weight averaging. The inference accuracy of the model on Stanford Car-196 is 93.68% without test-time augmentation or model ensemble. Hereafter referred to as EfficientNet-B3.

The $C_d = 0.2$ is experimentally verified in two common datasets by comparing the deletion process and visualization of samples.

Since model-agnostic methods assume that CNN models are black-box and the input image from the real world is not limited to ImageNet, the preprocess input instruction (ensure image colour channel zero-centred) is not applied to adjust the RGB channel distribution of the input image. The prediction results and saliency value distribution characteristics for the samples in this paper are shown in Table 1, where green indicates a confidence score greater than 60% or r^k_{dsm} greater than 1%; orange indicates that the results do not match Ground Truth (GT) or r^k_{dsm} is less than 0.2%.

Referring to the empirical results of multiple samples in Table 1, we conclude several descriptions as follows:

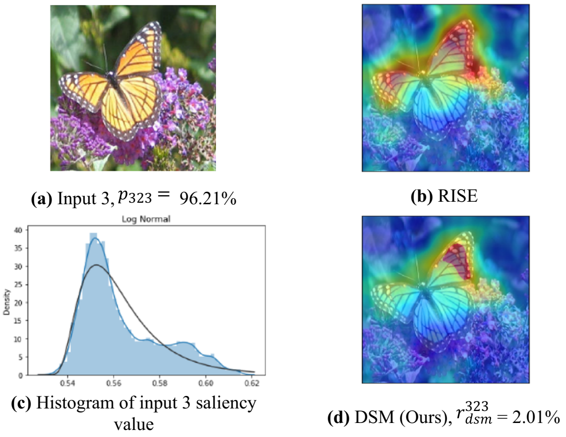
1. **High robustness** When the value of r^k_{dsm} is equal to or greater than 1%, it can be validated that the image has enough essential feature regions for the model to recognize. The inference process can robustly overcome the perturbation, even mostly covering the decisive salient region. All results for samples with r^k_{dsm} values above 1% in Table 1 are Ground Truth and Top 1 class, even if the score probability results are numerically low (e.g., sample 7).
2. **Barely acceptable** When r^k_{dsm} is between 0.2% and 1%, the robustness of model inference is relatively weak. Perturbation large enough to cover the decisive salient region still impacts the results.
3. **Poor robustness** When r^k_{dsm} is below 0.2% or far worse, the robustness deteriorates rapidly. The model inference is highly susceptible to negligible occlusion in the input image. Even if the occlusion is only 10 to 30 pixels of a 224 × 224 image, which is harmless for human perception, there is a high possibility of false prediction for CNN models.



(a) Input 2, $p_1 = 32.36\%$ **(b)** RISE

(c) Histogram of input 2 saliency value **(d)** DSM (Ours), $r^1_{dsm} = 0.32\%$

Fig. 3 Comparisons of the visual explanations on sample 2 by RISE and DSM using ResNet50



(a) Input 3, $p_{323} = 96.21\%$ **(b)** RISE

(c) Histogram of input 3 saliency value **(d)** DSM (Ours), $r^{323}_{dsm} = 2.01\%$

Fig. 4 Comparisons of the visual explanations on sample 3 by RISE and DSM using ResNet50

4.2 Experiments on ImageNet

When observing visual explanations subjectively using DSM, we can discover the feature regions' actual influence and avoid the confusion caused by subordinated features that have insufficient influences on inference. In Figs. 3 and 4, a comparison can be found between RISE and DSM's differences in saliency maps using ResNet50.

In Fig. 3a–b, the original visual explanation for sample 2 gives the impression that each goldfish is of equal importance for the model inference, i.e., the model focuses equally on the features represented by multiple goldfish. In contrast, our approach shows in Fig. 3d that the region represented by one and only one goldfish in the fish school has the maximum value and its proximity of the feature saliency, while the salient regions of other goldfish do not. Our visual explanation also better reflects the purpose for which the deletion process was set up. It is to discover regions with profound feature information which have a significant impact on the classification score, but with as few pixels as possible, through perturbation like masks [19].

Figure 4c shows that sample 3 has a lower absolute value of Kurtosis relative to sample 2 and a much higher mean saliency value in Fig. 3c. Combined with DSM shown in Fig. 4d, the visual explanation reflects that the focused feature area is adequate. The pixels with the maximum saliency value are concentrated on the right-wing instead of the relatively uniform distribution on both wings shown in Fig. 4b. r^k_{dsm} of sample 3 is quantified as in Table 1. It visualizes the discrepancy in the distribution of the saliency values. Sample 3 has a larger decisive salient region than sample 2, which is consistent with the high probability score.

To further verify the discriminative effect of DSM on the feature regions and the feasibility of robustness assessment, the salient region of the Top-1 class of the prediction is perturbed with the mask motivated by adversarial erasing [43] or by patch permutations, as in Fig. 5. The visual explanation of DSM guides the size and location of the mask. Patch permutations boost the model to learn features of different levels of granularity when training [44]. Simultaneously, it demonstrates the robustness of the model to spatial structural information disturbance when inferring [38]. Compared with occlusions and patch permutations, other natural and spatial perturbations, e.g., Gaussian blur, are tested to have relatively minute disturbance on the model inference.

The inference process with poor robustness fails for the perturbed input sample to get the correct prediction. Once the only goldfish representing the decisive saliency is occluded partially or perturbed by shuffle operation, as in Fig. 5a–d, the confidence score of the goldfish is reduced to lower than the probability of other classes, leading to misclassification. In contrast, input images with a sufficiently large coverage rate of DSM can maintain correct results and high scores, even after a severer perturbation or being shuffled into smaller grids than the previous sample, as in Fig. 5e–f. The result indicates enough high-importance feature areas for the model to recognize.

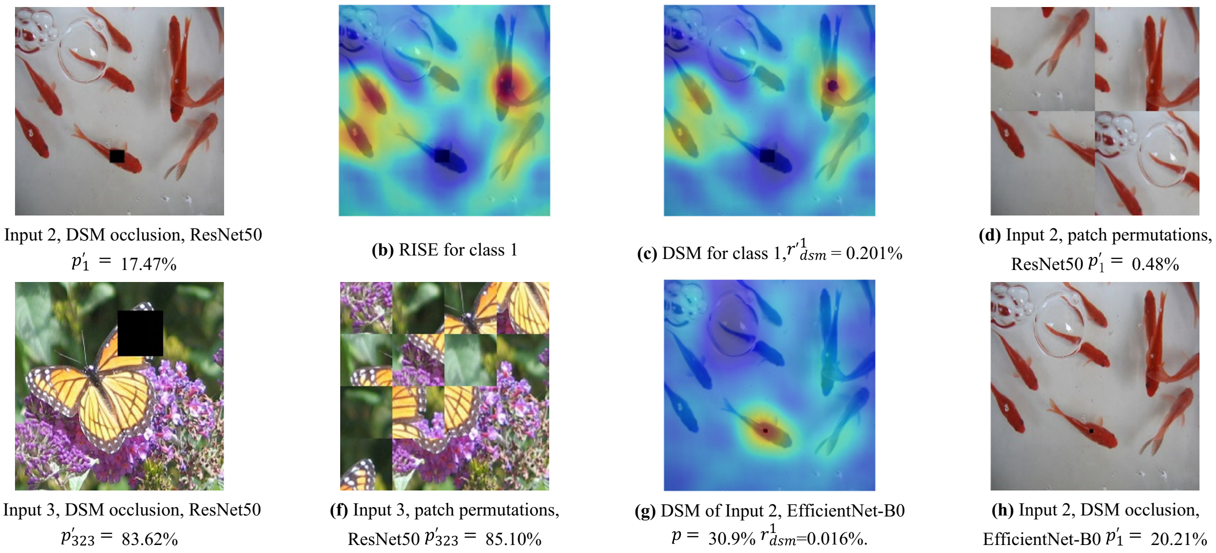
The samples above were further tested by comparing Decisive Saliency Maps of several CNN models with different network depths. In well-designed CNN models, the deeper the structure and higher the accuracy, the better the models can exploit the fine-grained features. Nevertheless, many samples preliminarily verify that when the features of interest are similar to different CNN models, overfitting may manifest in focusing excessively on limited fine-grained regions due to deeper layers, leading to poor robustness of models. Too much attention to too small features is not conducive to model generalization in real-world applications.

As shown in DSM, EfficientNet-B0 focuses on the same goldfish as ResNet50 in Fig. 5g, but the region of decisive saliency is much smaller. We perturb salient regions guided by DSM as in Fig. 5h and find that even a negligible perturbation to the human eye already caused the false prediction of EfficientNet-B0. Therefore, if the robustness of the model for real-world applications is a concern, a deeper and more complex model, while performing better, may not be the most appropriate choice without sufficient degraded samples and data augmentation.

DSM is also applicable to the class discriminative inferences of input samples containing objects of multiple classes. As in Fig. 6a, input sample 4 has two classes, bull mastiff and tiger cat, which are the Top-2 classification results by ResNet50. It can be found that a small amount of perturbation in the decisive salient region representing the tiger cat, shown in Fig. 6e, significantly reduces the confidence score for the cat as in Fig. 6f. Compared to the saliency map by RISE in Fig. 6d, DSM indicates clearly that only the cat's mouth and nose, not its head as a whole, are being focused on by the model. The confidence score of the first class (bull mastiff) increases remarkably after the salient region of the second class is disturbed with the indication of DSM, allowing the prediction to be very "confident" in Fig. 6f. When the features that ResNet50 and EfficientNet-B0 focus on are approximate, the inference process of EfficientNet-B0 is less robust in comparison again, as shown in Fig. 6k–l. The coverage rate of the DSM of both classes is fairly low. Occlusion as a tiny mask on decisive salient regions of the top class lowers its score to third in Fig. 6m. The input sample is shuffled into different levels of granularity in Fig. 6n–p. In most patch permutations cases, class 243 scores are higher than class 282, depending on the integrity of decisive salient regions after shuffle operation.

Figure 7 shows a comparison of more visual explanations of input samples. The visual explanations are generated by RISE, our method DSM, and the most representative model-specific method, GRAD-CAM. DSM improvement focuses on highlighting the most important features compared to other visual explanations, such as focusing on the animal's eyes in samples 5 and 8. For sample 7 bubbles which may represent the dispersion problem [29], DSM visualizes the bubble contours unambiguously while suppressing the noise.

For the misclassification cases, DSM can reflect the risk of untrustworthiness in the inference, and facilitate the detection of false predictions. The other two visual explanations are incapable of verification during interpretation. For puzzling sample 9, the valley, both CNN models without preprocessing misclassified it as the cliff, and they are highly susceptible to perturbation to misclassify input as the cliff dwelling. However, confidence scores greater than 60% cannot expose the risk of misclassification. But DSM reveals that the value of r^k_{dsm} is too small. The high-saliency region is unnoticeable via visualization, thus exposing misclassification risk.



(a) Input 2, DSM occlusion, ResNet50 $p'_1 = 17.47\%$

(b) RISE for class 1

(c) DSM for class 1,$r'^{1}_{dsm} = 0.201\%$

(d) Input 2, patch permutations, ResNet50 $p'_1 = 0.48\%$

(e) Input 3, DSM occlusion, ResNet50 $p'_{323} = 83.62\%$

(f) Input 3, patch permutations, ResNet50 $p'_{323} = 85.10\%$

(g) DSM of Input 2, EfficientNet-B0 $p = 30.9\%$ r^{1}_{dsm}=0.016%.

(h) Input 2, DSM occlusion, EfficientNet-B0 $p'_1 = 20.21\%$

Fig. 5 Comparisons of RISE and DSM using ResNet50 and EficientNet-B0. In (a)-(c), the GT class dropped from first to third after perturbation in decisive salient regions. In (d), the decisive salient region is perturbed by patch permutations into 2 × 2 grids. The prediction fails consequently. The inference with ResNet50 under large occlusion or permutated into 4 × 4 grids shows good robustness in (e)-(f). The confidence score is scarcely affected, and the value of r^{323}_{dsm} is big. Using EfficientNet-B0 for input sample 2, DSM shows in (g) that the value of r^{1}_{dsm} is too low (only 0.016%) and displays poor robustness. Therefore a negligible perturbation in (h) has reduced the score of the GT class from the first to the second

4.3 Experiments on stanford car-196

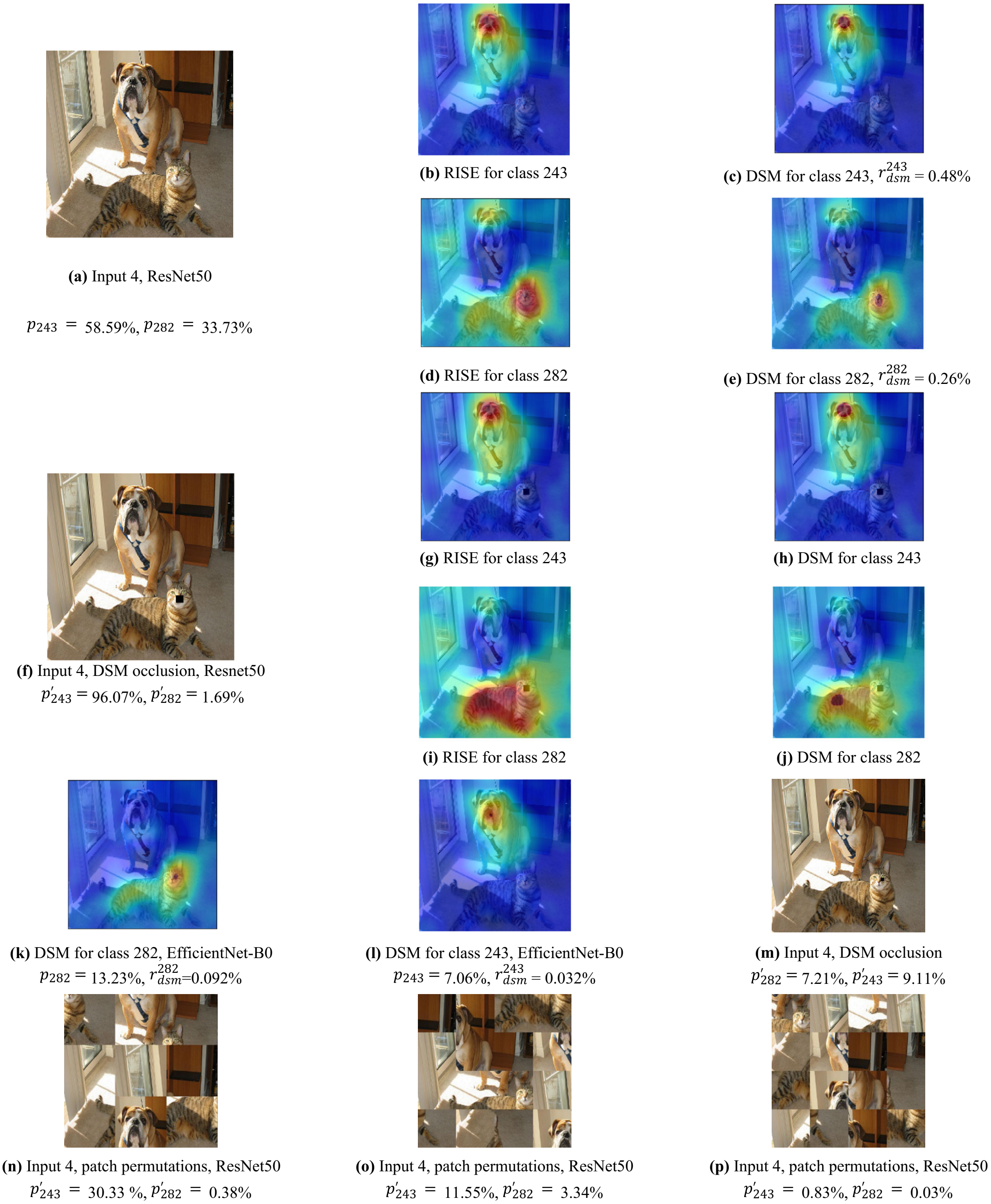
For encapsulated Deep Learning models in industrial applications or even traditional pattern-based machine vision algorithms, visual explanations are validated by RISE and DSM, which are perturbation-based approaches and suitable for black-box models. The only premise is that the industrial machine vision systems can output the results with probability scores correctly matching each of the massive perturbed input samples. With transfer learning and multiple regularizations, EfficientNet-B3 simulating industrial applications obtains higher accuracy and acceptable confidence scores (> 80%) due to better model performance and a smaller volume of class labels in the fine-grained dataset Stanford Car-196 compared to ImageNet.

When referring to the first and second columns of Fig. 8, it is difficult for end-users to detect the robustness risk based only on visual explanations and confidence scores. Provided that the predictions are correct, the difference in the size of essential salient regions produced by samples with high robustness of model inference and those without is difficult to be distinguished precisely. Calculating AUC using the deletion process is costly in GPU inference. However, the difference in AUC values still does not directly indicate the discrepancy in robustness, referring to the fourth column of Fig. 8.

Figure 8i shows the input sample with high robustness of model inference. The distribution histogram in Fig. 8y shows that the mean saliency value is high. The fit error between the Johnson unbounded and log-normal distribution is low. Whereas the sample with poor robustness of model inference is shown in Fig. 8m, it is seen that the mean saliency value is low as in Fig. 8z. The value of Kurtosis is much higher, and the distribution has a long tail effect.

Combined with physical objects in the real world, the visual explanations of DSM in the third column of Fig. 8 focus on discriminative features such as vehicle front mesh grilles, emblems, headlights, and taillights. Despite the distinctiveness of implementation methods, the findings are basically consistent with the description of [45]. DSM reduces the confusion of background and unneeded physical features on subjective cognition.

When decisive salient regions and coverage rates are considerably larger, models can still have correct predictions with unaffected confidence scores on samples even if large areas of the images are perturbed, as in Fig. 9a–b. This is the case for most test set samples after various data augmentation and training optimization of the fine-tuned model. In contrast, when the decisive salient regions are small but inferences have relatively high confidence scores, the predictions are susceptible to the perturbation in decisive salient regions in Fig. 9c–f. Corresponding to the real-world application, it is equivalent to the situation where minor damage to an auxiliary vehicle part causes CNN models to fail to recognize the vehicle type. Though EfficientNet-B3 has better performance, deeper layers, and more keen attention to



(a) Input 4, ResNet50

$p_{243} = 58.59\%, p_{282} = 33.73\%$

(b) RISE for class 243

(c) DSM for class 243, $r^{243}_{dsm} = 0.48\%$

(d) RISE for class 282

(e) DSM for class 282, $r^{282}_{dsm} = 0.26\%$

(f) Input 4, DSM occlusion, Resnet50

$p'_{243} = 96.07\%, p'_{282} = 1.69\%$

(g) RISE for class 243

(h) DSM for class 243

(i) RISE for class 282

(j) DSM for class 282

(k) DSM for class 282, EfficientNet-B0

$p_{282} = 13.23\%, r^{282}_{dsm}$=0.092%

(l) DSM for class 243, EfficientNet-B0

$p_{243} = 7.06\%, r^{243}_{dsm} = 0.032\%$

(m) Input 4, DSM occlusion

$p'_{282} = 7.21\%, p'_{243} = 9.11\%$

(n) Input 4, patch permutations, ResNet50

$p'_{243} = 30.33\ \%, p'_{282} = 0.38\%$

(o) Input 4, patch permutations, ResNet50

$p'_{243} = 11.55\%, p'_{282} = 3.34\%$

(p) Input 4, patch permutations, ResNet50

$p'_{243} = 0.83\%, p'_{282} = 0.03\%$

Fig. 6 DSM and robustness assessment of the prediction of input sample 4. The input sample **a** includes two ImageNet classes, 243 and 282. **b–e** show the comparison of RISE and DSM. The perturbed sample **f** for class 282, which refers to DSM **e**, shows that inference using ResNet50 shifts to focus entirely on class 243 after partial perturbation, essentially removing the attention to class 282. In DSM using EfficientNet-B0, class 282 is the first and 243 is the fourth. Poor robustness is visualized in **k** and **l**. **m** shows that negligible perturbation changes the class sequence of the result. **n–p** demonstrate the input sample shuffled into 3 × 3 and 4 × 4 grids. In most permutation cases, class 243 remains higher scores than class 282

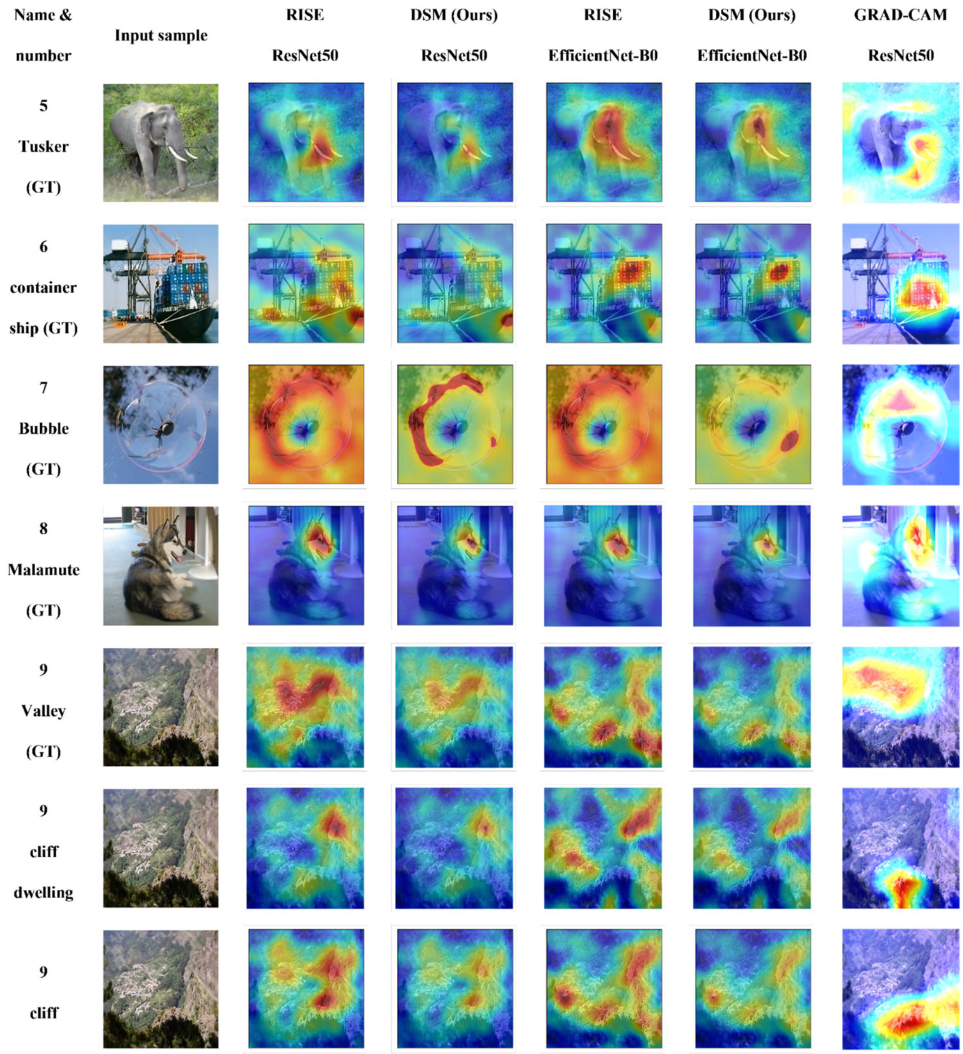


Fig. 7 Comparison of DSM and other visual explanations using Resnet50 and EfficientNet-b0 for more samples

discriminative features, its lack of robustness to infer certain classes or samples, e.g., images from the rearview of class 196 in the Stanford Cars dataset, is reflected by the utilization of DSM. DSM visually explains the variation of the prediction scores when in different granularity of patch permutations in Fig. 9g–l. The variation depends on the size of decisive salient regions and how the grids perturb the regions.

4.4 Class sensitivity evaluation

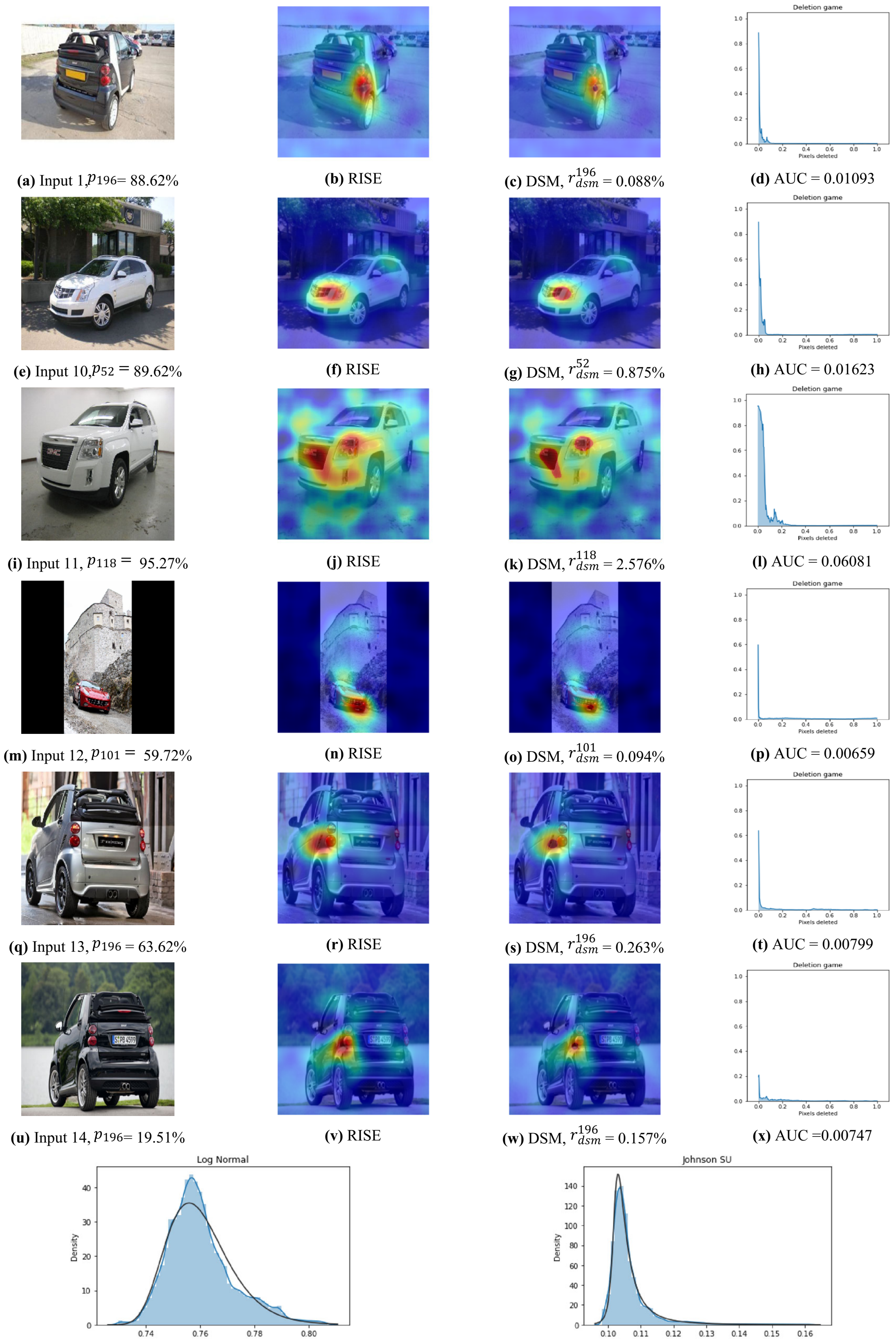
Class Sensitivity is defined and verified with different visual explanation methods in [29]. A responsible visual explanation in the image classification task should provide a different interpretation for each class. Besides, higher Class Sensitivity should display more discriminative and dissimilar visual explanations between saliency maps of classes with higher and lower scores. The advantages of DSM method regarding Class Sensitivity are presented in visual cognition and computation results of (dis)similarity metrics, as evaluated qualitatively and quantitatively.

Qualitative Evaluation Saliency maps of the lower-score classes provided by RISE are occasionally confusing. The visualizations tend to misguide the observers at first glance to convince enough, or even excessive discriminative features are considered during model inference.

Saliency maps by DSM are remarkably explicit and meaningful for lower-score classes, exposing that models could scarcely recognize correct features or salient regions.

The dissimilarity of saliency maps generated by RISE and DSM is shown in Fig. 10, comparing the classes with the highest and lowest scores. Optimization by DSM is prominent in subjective cognition.

Quantitative Evaluation Along with the Pearson Correlation Coefficient (CC), we apply several other commonly used similarity metrics for saliency maps generated by RISE and DSM. The dissimilarity between classes with the highest and lowest scores is calculated and compared.



(a) Input 1,p_{196}= 88.62% **(b)** RISE **(c)** DSM, $r_{dsm}^{196} = 0.088\%$ **(d)** AUC = 0.01093

(e) Input 10,p_{52} = 89.62% **(f)** RISE **(g)** DSM, $r_{dsm}^{52} = 0.875\%$ **(h)** AUC = 0.01623

(i) Input 11, p_{118} = 95.27% **(j)** RISE **(k)** DSM, $r_{dsm}^{118} = 2.576\%$ **(l)** AUC = 0.06081

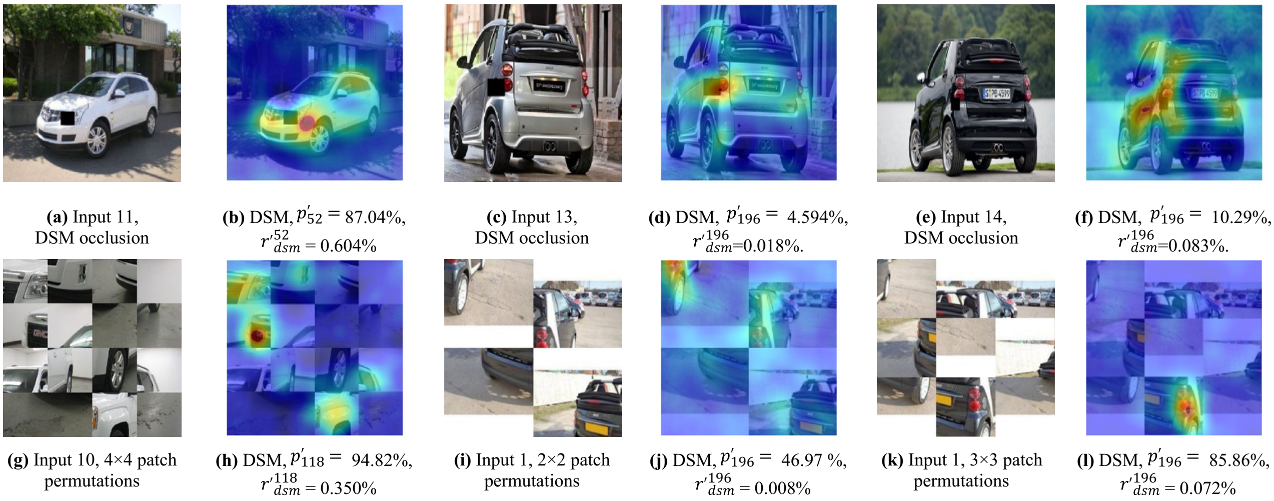
(m) Input 12, p_{101} = 59.72% **(n)** RISE **(o)** DSM, $r_{dsm}^{101} = 0.094\%$ **(p)** AUC = 0.00659

(q) Input 13, p_{196} = 63.62% **(r)** RISE **(s)** DSM, $r_{dsm}^{196} = 0.263\%$ **(t)** AUC = 0.00799

(u) Input 14, p_{196}= 19.51% **(v)** RISE **(w)** DSM, $r_{dsm}^{196} = 0.157\%$ **(x)** AUC =0.00747

(y) Histogram of saliency value of input 11 **(z)** Histogram of saliency value of input 12

Fig. 8 DSM and corresponding deletion AUC using EfficientNet-B3 model on samples from Stanford Car-196



(a) Input 11, DSM occlusion

(b) DSM, p'_{52} = 87.04%, r'^{52}_{dsm} = 0.604%

(c) Input 13, DSM occlusion

(d) DSM, p'_{196} = 4.594%, r'^{196}_{dsm}=0.018%.

(e) Input 14, DSM occlusion

(f) DSM, p'_{196} = 10.29%, r'^{196}_{dsm}=0.083%.

(g) Input 10, 4×4 patch permutations

(h) DSM, p'_{118} = 94.82%, r'^{118}_{dsm} = 0.350%

(i) Input 1, 2×2 patch permutations

(j) DSM, p'_{196} = 46.97 %, r'^{196}_{dsm} = 0.008%

(k) Input 1, 3×3 patch permutations

(l) DSM, p'_{196} = 85.86%, r'^{196}_{dsm} = 0.072%

Fig. 9 Decisive Saliency Maps for perturbed input samples with various robustness. The predictions of EfficientNet-B3 change accordingly. The scores do not drop in patch permutations simply due to more granular levels of grids. The variation of scores depends on the size of decisive salient regions and how the grids perturb the regions

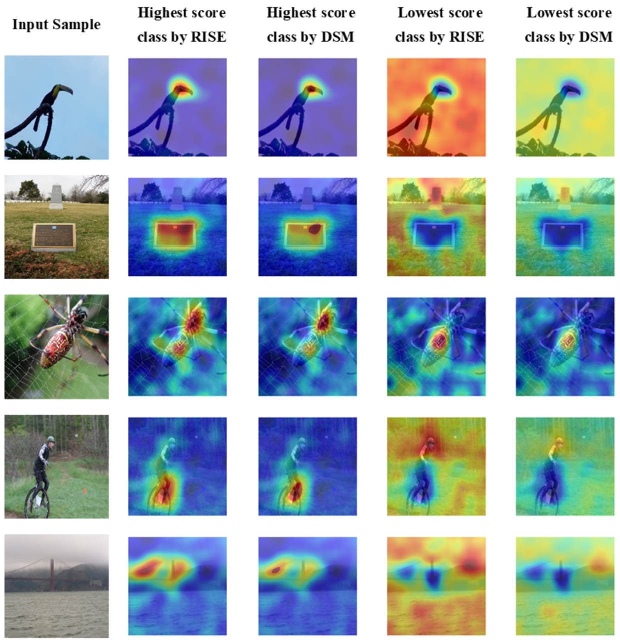

Input Sample
Highest score class by RISE
Highest score class by DSM
Lowest score class by RISE
Lowest score class by DSM

Fig. 10 The dissimilarity between saliency maps of the classes with the highest and lowest scores. Saliency maps of the lowest scores by DSM are remarkably meaningful

Table 2 The dissimilarity evaluation results between classes with highest and lowest scores from the aforementioned metrics. Smaller values are desired in dissimilarity evaluation

Method	SIM	NKL	CC	NSS
RISE	0.740	0.756	− 0.288	− 0.342
DSM	**0.737**	**0.752**	− 0.288	**− 0.350**

Bold values represent results whose metrics are better. Results without bold format mean equal performance

The Similarity (SIM) metric calculates the similarity index from the normalized saliency distributions of the predicted and ground truth saliency maps [46, 47]. The Kullback–Leibler divergence (KL) is a classical measure to estimate dissimilarities between the probability distribution of two maps, giving more penalty to false negatives. For better comparison, $NKL = 1 - KL$ is used in the evaluation [48]. The Normalized Scanpath Saliency (NSS) is an effective measurement sensitive to false positives and dissimilarity between prediction and ground truth [46, 47].

The saliency maps are normalized, respectively. Then top classes are set as ground truth in the calculation. We binarize the top-class saliency map in NSS with its mean saliency value as the threshold.

The pre-trained CNN model is ResNet50. The evaluation is conducted over a subset with more than 300 samples randomly picked from ImageNet. The sample amount approximates typical saliency benchmark datasets.

The results of CC are close to the experimental results in [29]. The symmetric computation of CC does not assume which saliency map is the ground truth. Thus, it cannot separate differences from false positives or false negatives. Positive NSS indicates a consistent correlation between saliency maps, and negative NSS indicates apparent dissimilarity. Considering saliency maps generated by numerous perturbations are distributed more pervasively than real human eye fixation, other computational values of similarity metrics are higher correspondingly.

As shown in Table 2, smaller values mean larger dissimilarity, representing higher Class Sensitivity. Most metrics demonstrate a certain level of optimization by DSM method.

The overall evaluation results indicate that DSM is an improved method for Class Sensitivity qualitatively and quantitatively, illustrating the dissimilarity between the highest and lowest classes with increased efficiency.

5 Conclusion and future work

This paper proposes an optimized visual explanation called Decisive Saliency Map applicable to black-box models for image classification tasks. DSM can quantitatively calculate the discrepancy of influence and size of different salient regions, also embody extra information on the distribution of saliency value in visualization. Its function of robustness assessment of the model inference process is validated on ImageNet and Stanford Car-196 datasets.

Further research will be conducted to eliminate the influence of randomness on the quantitative metrics. Simultaneously, we will continue to study the visual explanations of Deep Learning models to promote the utilization in other CNN vision tasks, including object detection, instance segmentation, etc. Endeavors will be made to reliable deployment and promotion of visual explanations in the manufacturing environment, analyzing the selection of backbone networks to balance accuracy and robustness requirements.

Appendix: Metrics to evaluate class sensitivity

The (dis)similarity metrics of saliency maps for evaluating Class Sensitivity are listed below.

Saliency maps and related explanations of the classes with the highest and lowest scores can be defined as:

$$c_{max}, c_{min} = \arg max f(I), \arg min f(I)$$

$$SM_{max}, SM_{min} = E(I, f)_{c_{max}}, E(I, f)_{c_{min}} \tag{11}$$

The saliency maps SM_{max}, SM_{min} are normalized as required in SIM, KL, and NSS calculations. Then top classes are set as ground truth in the calculation.

In KL computation, ϵ is a regularization constant, with the value of 2.2204e-16 in usual. We binarize the top-class saliency map as $SM_{max}{}_i^B$ in NSS with its mean saliency value as the threshold.

$$SIM = \sum_{x=1}^{X} min(SM_{min}, SM_{max}) \tag{12}$$

$$KL = \sum_{x=1}^{X} SM_{min} * \log\left(\frac{SM_{min}}{SM_{max} + \epsilon} + \epsilon\right)$$

$$NKL = 1 - KL \tag{13}$$

$$CC = \frac{\mathrm{cov}(SM_{min}, SM_{max})}{\sigma_{SM_{min}} * \sigma_{SM_{max}}} \tag{14}$$

$$NSS\left(SM_{min}, SM_{max}{}_i^B\right) = \frac{1}{N}\sum_i SM_{min} \times SM_{max}{}_i^B$$

where

$$N = \sum_i SM_{max}{}_i^B \tag{15}$$

References

1. Yang, T., Zhang, T., Huang, L.: Detection of defects in voltage-dependent resistors using stacked-block-based convolutional neural networks. Vis. Comput. **37**, 1559–1567 (2021). https://doi.org/10.1007/s00371-020-01901-w
2. Patel, N., Mukherjee, S., Ying, L.: EREL-Net: A remedy for industrial bottle defect detection. International Conference on Software Maintenance. Lecture Notes in Computer Science, vol 11010. Springer, Cham. (2018). https://doi.org/10.1007/978-3-030-04375-9_39
3. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: a survey of case studies. NeurIPS: ML Retrospectives, Surveys & Meta-Analyses (2020). https://doi.org/10.1145/3533378
4. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Magazine, vol. 40, no. 2 (2019). https://doi.org/10.1609/aimag.v40i2.2850
5. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
6. Artificial Intelligence (AI) - Assessment of the robustness of neural networks. ISO/IEC Technical Report 24029–1:2021 (2021)
7. Martin, D., Heinzel, S., Von Bischhoffshausen, J. Kunze, Kühl, N.: Deep learning strategies for industrial surface defect detection systems. In: the Annual Hawaii International Conference on System Sciences (2022). https://doi.org/10.24251/hicss.2022.146
8. Vermeire, T., Laugel, T., Renard, X., Martens, D., Detyniecki, M.: How to choose an explainability method? Towards a methodical implementation of XAI in practice. Communications in Computer and Information Science, (2021). https://doi.org/10.1007/978-3-030-93736-2_39
9. Brundage, M. et al.: Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv: 2004.07213v2 (2020)
10. Wagner, J., Köhler, J. M., Gindele, T., Hetzel, L., Wiedemer, J. T., Behnke, S.: Interpretable and fine-grained visual explanations for convolutional neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9089–9099 (2019). https://doi.org/10.1109/CVPR.2019.00931
11. Ji, S., Li, J., Du, T., Li, B.: A survey on techniques, applications and security of machine learning interpretability. J. Comput. Res. Develop. **56**(10), 2071–2096 (2019)

12. Khorram, S., Lawson, T., Li, F.: iGOS++: integrated gradient optimized saliency by bilateral perturbations. CHIL '21: Proceedings of the Conference on Health, Inference, and Learning April, Pages 174–182. (2021). https://doi.org/10.1145/3450439.3451865
13. Finale, D., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608v2 (2017)
14. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
15. Zeiler, M. D., Fergus, R.: Visualizing and understanding convolutional networks. In European conference on computer vision, pp. 818–833. Springer (2014)
16. Simonyan, K., Vedaldi, A., Zisserman A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
17. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
19. Fong, R. C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3449–3457 (2017). https://doi.org/10.1109/ICCV.2017.371
20. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (2018)
21. Ribeiro, M. T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: AAAI Conference on Artificial Intelligence, pp 1527–1535 (2018)
22. Barredo Arrieta, A., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform. Fus. **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319
24. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
25. Wang, H. et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 111–119, (2020). https://doi.org/10.1109/CVPRW50498.2020.00020.
26. Cheng, K., Wang, N., Shi, W., Zhan, Y.: Research advances in the interpretability of deep learning. J. Comput. Res. Develop. **57**, 1208 (2020). https://doi.org/10.7544/ISSN1000-1239.2020.20190485
27. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018). https://doi.org/10.1145/3236009
28. Fong, R., Patrick, M., Vedaldi A.: Understanding deep networks via extremal perturbations and smooth masks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2950–2958 (2019). https://doi.org/10.1109/ICCV.2019.00304
29. Li, X., Shi, Y., Li, H., Bai, W., Song, Y., Cao, C., Chen, L.: An experimental study of quantitative evaluations on saliency methods. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 3200–3208 (2021). https://doi.org/10.1145/3447548.3467148
30. Keller, P.R., Keller, M.M.: Visual cues: practical data visualization. IEEE Computer Society Press, Los Alamitos (1993)
31. Chen, W., Zhang, S., Lu, A., Zhao, Y.: Guide for Data Visualization (In Chinese). High Education Press (2020)
32. Johnson, N.L.: Systems of frequency curves generated by methods of translation. Biometrika **36**(1/2), 149 (1949). https://doi.org/10.2307/2332539
33. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013). https://doi.org/10.1109/ICCVW.2013.77
34. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proc. of International Conference on Machine Learning, pp. 6105–6114 (2019)
35. Morales, D.A., Talavera, E., Remeseiro, B.: Playing to distraction: towards a robust training of cnn classifiers through visual explanation techniques. Neural Comput. Appl. (2020). https://doi.org/10.1007/s00521-021-06282-2
36. Koffka, K.: Principles of Gestalt psychology. Routledge, Taylor & Francis Group, London (2013)
37. Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, 70:1321–1330 (2017)
38. Naseer, M., Ranasinghe, K., et al.: Intriguing properties of vision transformers. Neural Inform. Process. Syst. (NeurIPS 2021) **34**, 23296–23308 (2021)
39. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 13001–13008 (2020). https://doi.org/10.1609/aaai.v34i07.7000
40. Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., Choe, J.: CutMix: regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6022–6031 (2019). https://doi.org/10.1109/ICCV.2019.00612
41. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
43. Bargal, S.A., et al.: Guided zoom: zooming into network evidence to refine fine-grained model decisions. IEEE Transactions Pattern Anal. Mach. Intell. **43**(11), 4196–4202 (2021). https://doi.org/10.1109/TPAMI.2021.3054303
44. Du, R. et al.: Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol 12365. Springer, Cham. (2020). https://doi.org/10.1007/978-3-030-58565-5_10
45. Pei, H., Guo, R., Tan, Z., et al.: Fine-grained classification of automobile front face modeling based on Gestalt psychology. Vis. Comput. (2022). https://doi.org/10.1007/s00371-022-02506-1
46. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models. IEEE Transactions Pattern Anal. Mach. Intell. **41**(3), 740–757 (2019). https://doi.org/10.1109/TPAMI.2018.2815601
47. Riche N, Duvinage M, Mancas M, Gosselin B, Dutoit T.: Saliency and human fixations: state-of-the-art and study of comparison metrics. In IEEE International Conference on Computer Vision, pp. 1153–1160 (2013). https://doi.org/10.1109/ICCV.2013.147
48. Emami, M., Hoberock, L.L.: Selection of a best metric and evaluation of bottom-up visual saliency models. Image Vis. Comput.

31(10), 796–808 (2013). https://doi.org/10.1016/j.imavis.2013.08.004

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xiaoshun Xu received the B.S. degree in Communication Engineering from Tongji University, Shanghai, China, in 2007, and the M.E. degree in Communication Engineering from Fudan University, Shanghai, in 2011. He is currently pursuing a Ph.D. degree in Advanced Manufacturing at Shanghai Jiao Tong University, and working as the senior intelligent equipment manager at SAIC General Motors Corporation Limited, Shanghai. His research interests include Computer Vision and Intelligent Manufacturing.

Jinqiu Mo received the B.S. degree and Ph.D. degree in Mechanical Manufacturing and Automation form Zhejiang University, China, in 1991 and 1997, respectively. She is currently an associate professor at school of mechanical engineering, Shanghai Jiao Tong University. Her research interests include design and coordinated control of intelligent and precision electromechanical system, Computer Vision.