Randomized Channel-pass Mask for Channel-wise Explanation of Black-box Models

Hirotaka Hachiya[1] and Daiki Nisawa[1]

Wakayama University, 930 Sakaedan, Wakayama, Wakayama, 640-8510, Japan
hhachiya@wakayama-u.ac.jp

Abstract. In recent years, there has been active research on interpreting the classification results of deep models. Among these methods, MC-RISE enables pixel-color-wise interpretation based on the model output for images where pixels have been randomly replaced with a predetermined color. However, this approach requires manually preparing the appropriate color candidates. This study proposes a pixel-channel-wise interpretation method using a Randomized Channel-pass Mask (RaCM), which directly evaluates the importance of the original RGB values of an image through randomly generated masks that pass or exclude color channels of each pixel. Experiments are conducted using the German Traffic Sign Recognition Benchmark and ImageNet datasets. The effectiveness of the proposed method is demonstrated through evaluation metrics such as Insertion, Deletion, and Average DCC.

Keywords: Model interpretation · Black box · Visualization

1 Introduction

As machine learning is widely applied to image classification, the demand for methods that explain and visualize classification results is rapidly increasing. By identifying which parts of the image the model focuses on for classification, it becomes possible to interpret the model's behavior.

Model interpretation methods can be divided into two categories: white-box methods and black-box methods. White-box methods attempt to understand how the model makes classification decisions based on the internal structure and parameter values of the model. For example, these methods visualize which features influence the classification by utilizing feature maps which are the outputs of each layer of the model, e.g., convolution and gradient information of specific layers. Examples of such interpretation methods include Gradient-weighted Class Activation Mapping (Grad-CAM) [10] and Group score-weighted Class Activation Mapping (Group-CAM) [13].

On the other hand, black-box methods aim to understand the model's behavior from the outside without using internal information of the model. These methods are useful when the model is complex and its internal structure is opaque. Black-box methods make localized changes to the image, such as masking, and evaluate how these changes affect the model's predictions to visualize

which features influence the classification. Examples of these visualization methods include Local Interpretable Model-agnostic Explanations (LIME) [8] and Randomized Input Sampling for Explanation of Black-box Models (RISE) [6].

However, while these methods can visualize which locations affect the classification, they cannot show which color information influences it. Typically, the influence of a particular color in the detected region is judged based on human intuition. However, when image channels are not in the standard RGB format, for instance using HSV, or when dealing with image sets like 2D observations of precipitation, pressure, and temperature in weather forecasting, human intuition alone may not be sufficient for interpretation.

To overcome this problem, Multi-Color RISE (MC-RISE) [3] has been proposed to visualize color-specific influences. MC-RISE interprets which colors influence the classification based on the model's output for images with randomly replaced pixels with a predetermined color. However, MC-RISE requires manually setting color candidates in advance, making evaluating the importance of a diverse range of colors difficult.

In this study, we propose a black-box model interpretation method called Randomized Channel-pass Mask (RaCM). This method generates random channel-pass masks that pass or exclude the color channels of each pixel, enabling direct evaluation of the importance of the image's original RGB values. By analyzing the model output for these masked images, RaCM can determine the significance of each color channel more efficiently and accurately. Experiments are conducted using two datasets: the German Traffic Sign Recognition Benchmark (GTSRB) and ImageNet. The effectiveness of the proposed method is demonstrated using evaluation metrics such as Insertion, Deletion [6], and average DCC [7].

The main contributions of this paper are summarized as follows:

1. We propose RaCM, a novel black-box model interpretation method that directly evaluates the importance of the original RGB values at each pixel more efficiently and accurately than existing methods like MC-RISE. RaCM achieves this through randomly generated channel-pass masks that pass or exclude the color channels of each pixel.
2. We conducted extensive experiments on the GTSRB and ImageNet datasets, demonstrating that RaCM outperforms existing methods in terms of visualization performance and computational efficiency.

After this introductory section, the remainder of this paper is organized as follows. Section 2 reviews related works, and Section 3 details the proposed method. Section 4 describes the experimental evaluation and discussion, and the conclusion is presented in Section 5.

2 Related works

This section reviews the formulation and related work of model interpretation methods.

2.1 Formulation

Let $I \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times N^{\mathrm{ch}}}$ denote an input image, where N^{w}, N^{h}, and N^{ch} are the width, height, and number of channels, respectively. Let $M \in \{0,1\}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times N^{\mathrm{mask}}}$ be a random binary mask, where N^{mask} is the number of masks. Let $\sigma^k(I) \in \mathbb{R}$ represent the class score for the k-th class predicted by a deep model $\sigma(\cdot)$ given the image I. Additionally, let $\mathcal{I}_k \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times N^{\mathrm{ch}}}$ denote the importance (saliency) map, showing how important each pixel is for the model's prediction of the k-class score $\sigma^k(I)$. Finally, let $X(p)$ and $X(p,c)$ indicate the values in pixel p and channel c of an image or mask of X, such as the input image I, mask M, and importance map $\mathcal{I}$.

2.2 Single-Channel Visualization Methods

We discuss representative white-box interpretation methods for deep models, such as Grad-CAM [10]. Grad-CAM is an interpretation method that highlights feature maps on an input image I using gradient information with respect to the model's output. In Grad-CAM, the average-pooled gradient is used as the importance of a feature map A^c, and an importance map $\mathcal{I}_k^{\mathrm{CAM}} \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times 1}$ is created by the weighted sum of the feature maps as follows:

$$\mathcal{I}_k^{\mathrm{CAM}} = \mathrm{upsample}\left(\mathrm{ReLU}\left(\sum_{c=1} \mathrm{avgPool}\left(\frac{\partial \sigma^k(I)}{\partial A^c}\right) A^c\right)\right), \tag{1}$$

where upsample$(\cdot)$ is to enlarge the feature map to the same size as the input image and avgPool$(\cdot)$ is to perform average pooling. However, since high-importance feature maps are not always related to the target class k in all locations, noise can be captured.

To address this issue, several extensions have been proposed: Grad-CAM++ [1] smooths gradients using second-order derivatives for better localization, Score-CAM [12] evaluates each channel's feature map based on the model's prediction without using gradients, Ablation-CAM [2] systematically removes portions of feature maps and evaluates the difference in the model's prediction, and Group-CAM [13] divides feature maps into groups and generates weights for each group. However, since these methods calculate importance maps using gradients and feature maps, it is difficult to visualize which parts contribute at the color channel level of the input image because information from multiple channels is mixed due to convolution operations.

On the other hand, RISE [6], a representative black-box method, evaluates how changes in the input image I affect the model's predictions $\sigma^k(\cdot)$. Specifically, RISE randomly generates masks M covering some features in the input image and evaluates whether it increases or decreases $\sigma^k(\cdot)$. This process is repeated with multiple masks, allowing highlighting the importance of features on the input image without internal information of the model such as gradients and feature maps.

2.3 Multi-Color Visualization Methods

To visualize the importance of color information in addition to pixels, MC-RISE [3] prepares color candidates $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N^{\mathrm{col}}} \in \mathbb{Z}^3$. It generates a low-resolution mask $M^{\mathrm{low}}(p, i)$ by randomly selecting a pixel p and color $\boldsymbol{c}_i$ as follows:

$$M^{\mathrm{low}}(p,i) = \begin{cases} 1 & \text{if a pixel } p \text{ and a color } \boldsymbol{c}_i \text{ is selected,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

By enlarging M^{low} to the same resolution as the input image I using bilinear interpolation, a smoothed random pixel-color mask $M^{\mathrm{MC}} \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times N^{\mathrm{col}}}$ is created. Meanwhile, a non-targeted mask $M^{\mathrm{MC}_0}(p) = 1 - \delta\left(\sum_{i=1}^{N^{\mathrm{col}}} M^{\mathrm{MC}}(p,i) > 0\right) \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times 1}$ is calculated to indicate whether none of color candidate is selected at pixel p, where $\delta(x)$ is an indicator function that takes 1 if x is true and 0 otherwise. Then, the input image is altered by applying two masks as follows:

$$I^{\mathrm{MC}}(p; M^{\mathrm{MC}}) = I(p) M^{\mathrm{MC}_0}(p) + \sum_{i=1}^{N^{\mathrm{col}}} \boldsymbol{c}_i M^{\mathrm{MC}}(p,i). \tag{3}$$

Finally, the importance map $\mathcal{I}_k^{\mathrm{MC}}$ is obtained by the difference between the model's predictions $\sigma^k(\cdot)$ when a pixel p is overlaid with a color $\boldsymbol{c}_i$ (i.e., $M^{\mathrm{MC}}(p,i) = 1$), and when the pixel is left unchanged (i.e., $M^{\mathrm{MC}_0}(p) = 1$), as follows:

$$\begin{aligned} \mathcal{I}_k^{\mathrm{MC}}(p,i) = {} & \mathbb{E}_{M^{\mathrm{MC}}}\left[\sigma^k(I^{\mathrm{MC}}(p; M^{\mathrm{MC}}) \,\middle|\, M^{\mathrm{MC}}(p,i) = 1\right] \\ & - \mathbb{E}_{M^{\mathrm{MC}}}\left[\sigma^k(I^{\mathrm{MC}}(p; M^{\mathrm{MC}}) \,\middle|\, M^{\mathrm{MC}_0}(p) = 1\right], \end{aligned} \tag{4}$$

where the second term corresponds to the bias term, indicating whether the importance map's sign is positive or negative.

2.4 Analysis of MC-RISE

Since a color is uniformly selected from N^{col} candidates, the probability that pixel p and color i are selected is $P[M^{\mathrm{MC}}(p,i) = 1] = \frac{P^{\mathrm{pix}}}{N^{\mathrm{col}}}$ where P^{pix} is the probability of individually selecting a pixel p, and the probability that pixel p is not selected is $P[M^{\mathrm{MC}_0}(p) = 1] = 1 - P^{\mathrm{pix}}$. Therefore, expanding Eq. 4 and estimating the importance map $\mathcal{I}_k^{\mathrm{MC}}$ using Monte Carlo sampling results in the following:

$$\mathcal{I}_k^{\mathrm{MC}}(p,i) \approx \frac{1}{N^{\mathrm{mask}}} \sum_{m=1} \left(\frac{N^{\mathrm{col}} M_m^{\mathrm{MC}}(p,i)}{P^{\mathrm{pix}}} - \frac{M_m^{\mathrm{MC}_0}(p)}{1 - P^{\mathrm{pix}}} \right) \sigma^k\left(I^{\mathrm{MC}}(p; M_m^{\mathrm{MC}})\right), \tag{5}$$

where N^{mask} is the number of randomly generated masks. MC-RISE can visualize the importance of a pixel p and a predefined color $\boldsymbol{c}_i$ if there are many enough masks.

However, if there are different colored parts within a single object of target class k, such as a bird and chameleon, or if there are multiple target objects with different colors to be visualized, it becomes significantly challenging to prepare appropriate color candidates that include all the important colors in advance. Additionally, since MC-RISE evaluates the importance by randomly overlaying a color candidate $\boldsymbol{c}_i$ on the input image I as described in Eq. 4, not only the color but also the shape of the target object can be changed. For instance, if the color dot is placed near the borders of the objects, it may introduce considerable change, leading to inefficient evaluation of the importance.

3 Proposed Method

To address the issues of MC-RISE, this study proposes a Randomized Channel-pass Mask (RaCM) to directly evaluate the importance of the color information at each pixel using the original RGB values from the input image.

3.1 Random Channel-pass Mask

Instead of preparing predefined color candidates, RaCM generates random masks passing or excluding color channels c of each pixel p from the input image I, enabling efficient evaluation of the importance of color information. A channel-pass mask $M^{\mathrm{CM}} \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times N^{\mathrm{ch}}}$ is randomly created as follows:

$$M^{\mathrm{CM}}(p,c) = \begin{cases} 1 & \text{if a pixel } p \text{ and channel } c \text{ is selected to be passed,} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Next, a bias $M^{\mathrm{CM}_0} \in \mathbb{R}^{N^{\mathrm{h}} \times N^{\mathrm{w}} \times 1}$ is created to remove the bias effect from the mask, as follows:

$$M^{\mathrm{CM}_0}(p) = 1 - \delta\left(\sum_{c=1}^{N^{\mathrm{ch}}} M^{\mathrm{CM}}(p,c) > 0\right). \tag{7}$$

Then, the mask is applied to the input image I as follows:

$$I^{\mathrm{CM}}(p; M^{\mathrm{CM}}) = \sum_{c=1}^{N^{\mathrm{ch}}} I(p) M^{\mathrm{CM}}(p,c). \tag{8}$$

Fig. 1 depicts an example of the mask M^{CM} and the masked image I^{CM}. In the mask M^{CM}, pixels in black indicate excluding all color channels, while pixels in other colors indicate that the corresponding color channels are passed. For example, the red and green channels remain in the yellow regions.

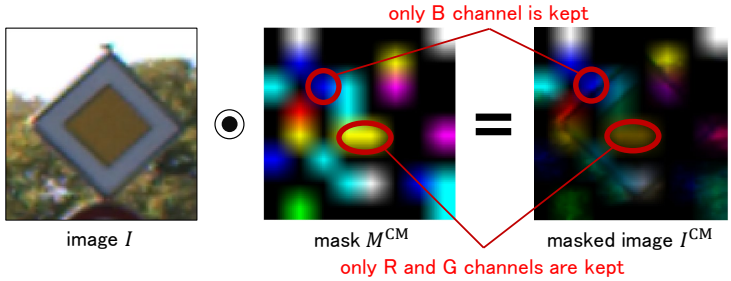

Fig. 1: An example of a channel-pass mask, M^{CM}, retains randomly selected pixels p, and channels c, resulting in a masked image I^{CM}. Pixels in black indicate excluding all color channels, while pixels in other colors, such as blue, yellow, cyan, and magenta, indicate that the corresponding color channels are passed.

3.2 Pixel-channel-wise importance map

The importance map $\mathcal{I}_k^{\mathrm{CM}}$ is obtained by the difference between the model's predictions $\sigma^k(\cdot)$ when a channel c is passed at a pixel p (i.e., $M^{\mathrm{CM}}(p,c)=1$), and when all channels are excluded (i.e., $M^{\mathrm{CM}_0}(p)=1$), as follows:

$$\begin{aligned}\mathcal{I}_k^{\mathrm{CM}}(p,c) = &\mathbb{E}_{M^{\mathrm{CM}}}\left[\sigma_k\left(I^{\mathrm{CM}}(p;M^{\mathrm{CM}})\right) \,\middle|\, M^{\mathrm{CM}}(p,c)=1\right] \\ &- \mathbb{E}_{M^{\mathrm{CM}}}\left[\sigma_k\left(I^{\mathrm{CM}}(p;M^{\mathrm{CM}})\right) \,\middle|\, M^{\mathrm{CM}_0}(p)=1\right], \end{aligned} \tag{9}$$

where the second term corresponds to the bias term, indicating whether the importance map's sign is positive or negative.

Since a channel is uniformly selected at a randomly selected pixel p, the probability that pixel p and channel c are selected to be passed is $P[M^{\mathrm{CM}}(p,c)=1]=\frac{2^{N^{\mathrm{ch}}-1}\times P^{\mathrm{pix}}}{2^{N^{\mathrm{ch}}}-1}$ and the probability that pixel p is not selected is $P[M^{\mathrm{CM}_0}(p)=1]=1-P^{\mathrm{pix}}$. Therefore, we can expand Eq. 9 and estimate the importance map $\mathcal{I}_k^{\mathrm{CM}}$ using Monte Carlo sampling for masks as follows:

$$\begin{aligned}&\mathcal{I}_k^{\mathrm{CM}}(p,c) \\ &\approx \frac{1}{N^{\mathrm{mask}}}\sum_{m=1}\left(\frac{2^{N^{\mathrm{ch}}-1}\times M_m^{\mathrm{CM}}(p,c)}{(2^{N^{\mathrm{ch}}}-1)P^{\mathrm{pix}}} - \frac{M_m^{\mathrm{CM}_0}(p)}{1-P^{\mathrm{pix}}}\right)\sigma^k\left(I^{\mathrm{CM}}(p;M_m^{\mathrm{CM}})\right)\end{aligned} \tag{10}$$

The detailed derivation is described in the appendix.

3.3 Flow of RaCM

Fig. 2 depicts the flow of the proposed RaCM method for a black-box model interpretation. For an input image I, N^{mask} channel-pass masks M^{CM} are randomly generated. Then, N^{mask} masked images I^{CM} are created by applying these masks (in Eq. 8) to the input image I and inputted to a deep model $\sigma(\cdot)$ to calculate corresponding predictions. If the class score is high, the mask retains useful

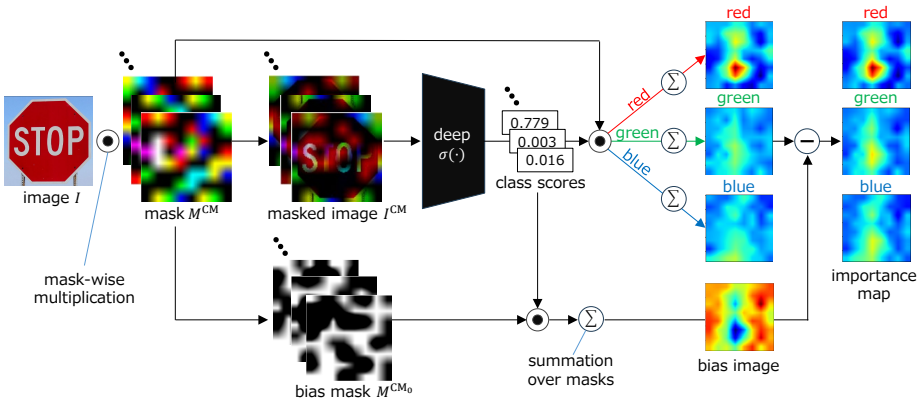

Fig. 2: Flow of the proposed RaCM method for a black-box model interpretation.

pixels p and channels c for classification, and if the class score is low, the useful information is considered not retained. The final importance map $\mathcal{I}^{\mathrm{CM}}$ is outputted by subtracting the bias image from the weighted sum of the masks M^{CM} (in Eq. 16) using the class scores as weighting factors for each RGB channel.

3.4 Analysis of MC-RISE vs. RaCM

To explain the difference between MC-RIS and RaCM, let us focus on the probability of masks used to estimate the importance maps (Eqs. 5 and 16). Assuming $N^{\mathrm{col}} = 7$ color candidates (red, green, blue, yellow, magenta, cyan, and white) to evaluate the possible combinations of RGB in MC-RISE, the mask probability for MC-RISE is $P(M^{\mathrm{MC}}(p,i) = 1) = \frac{P^{\mathrm{pix}}}{7}$, while the mask probability for RaCM is $P(M^{\mathrm{CM}}(p,c) = 1) = \frac{4P^{\mathrm{pix}}}{7}$ (see Eq. 15 in appendix), which is almost four times as high. This means that MC-RISE evaluates the RGB colors by replacing the color of each pixel with a randomly selected color candidate $\boldsymbol{c}_i$, while RaCM uses the color information of the image itself for each pixel, leading to more efficient evaluation of the RGB channels. Therefore, RaCM is expected to evaluate channels with fewer masks.

Moreover, in MC-RISE, the color information of the input image I remains in the masked image I^{MC} (in Eq. 3) only for non-targeted regions ($M^{\mathrm{MC}_0} = 1$). Therefore, the classification score for the mask depends on both the random pixel-color mask M^{MC} and non-targeted mask M^{MC_0}, and if important objects remain in the non-targeted masks, the mask may not be evaluated correctly. Specifically, as shown in Fig. 3, when the object (chameleon) is not targeted, the masked image I^{MC} leaves the object intact, so the targeted objects around the chameleon are evaluated as important. Additionally, as shown in Fig. 4, even when the target object is masked, MC-RISE overwrites the target object with the selected color $\boldsymbol{c}_i$, possibly changing its shape and lowering the classification score.

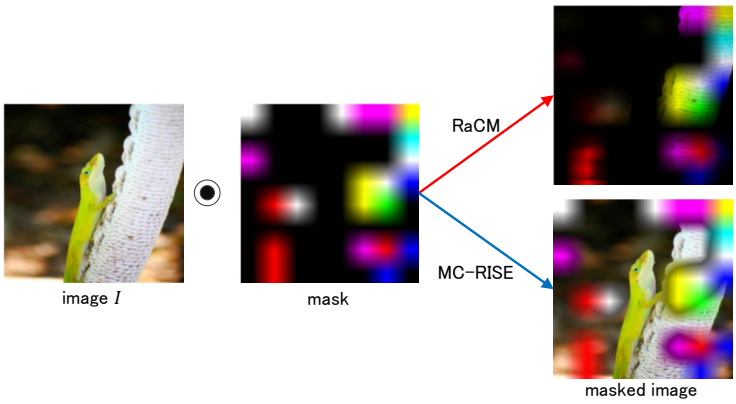

Fig. 3: An example of masked image when the target object is not masked

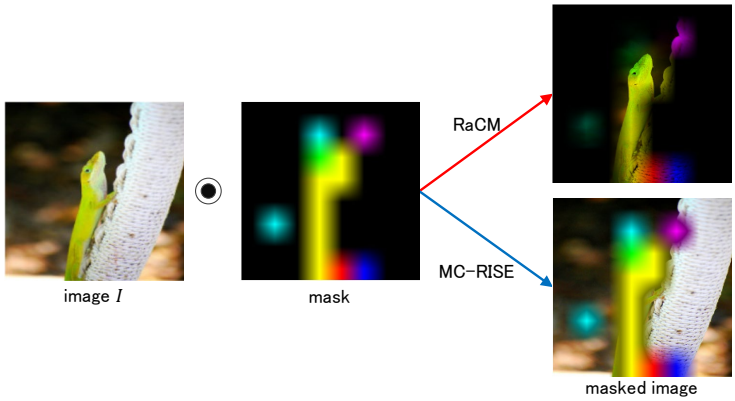

Fig. 4: An example of masked image when the target object is masked

In contrast, RaCM discards all RGB channels for non-targeted masks M^{CM_0}, so the target object does not remain as shown in Fig. 3. Moreover, the mask pixels M^{CM} retain the color channels of the image, allowing the target object to remain without changing its shape (although the color may change). Therefore, the importance of pixels and color channels can be evaluated more accurately.

3.5 Region Refinement using Grad-CAM

While RaCM can efficiently sample channels c, it is not efficient to sample possible combinations of pixels p from the entire image $N^{\mathrm{h}} \times N^{\mathrm{w}}$. Therefore, white-box methods are used to exclude regions that are obviously not important, such as the background, from the sampling target in advance. For example, when using Grad-CAM, the sampling target is limited to regions where the importance map $\mathcal{I}^{\mathrm{CAM}}$ (in Eq. 1) is greater than a threshold θ as follows:

$$M^{\mathrm{CM}\prime} = M^{\mathrm{CM}} \odot \delta\left(\mathcal{I}^{\mathrm{CAM}} > \theta\right), \tag{11}$$

where the threshold θ is set to a percentile value (e.g., 30%) of $\mathcal{I}^{\mathrm{CAM}}$ for each image.

4 Experiments

We demonstrate the effectiveness of the proposed RaCM method through experiments using the public datasets GTSRB (German Traffic Sign Recognition Benchmark) [5] and ImageNet [9].

4.1 Experimental setup

The model interpretation performance of MC-RISE [3], the proposed RaCM, and the proposed RaCM+Grad-CAM are evaluated. To evaluate the efficiency of the visualization, three different numbers of masks $N^{\mathrm{mask}} \in \{1000, 3000, 5000\}$ are prepared. The probability that each pixel is individually selected is set to $P^{\mathrm{pix}} = 0.5$. The color candidates for MC-RISE are the same seven colors (red, green, blue, yellow, magenta, cyan, and white) used in the proposed RaCM (see Section 3.2). The importance of the RGB channels is determined by the weighted average of the importance $\mathcal{I}_k^{\mathrm{MC}}$ for the color candidates obtained by MC-RISE as follows:

$$\mathcal{I}_k^{\mathrm{MC}}(p, c) = \frac{1}{4} \sum_{i=1}^{7} \boldsymbol{c}_i(c) \mathcal{I}_k^{\mathrm{MC}}(p, i), \tag{12}$$

where $\boldsymbol{c}_i(c)$ indicates the c-th element of a vector $\boldsymbol{c}_i$.

4.2 Evaluation Metrics

The standard performance evaluation metrics for a model interpretation methods are Insertion, Deletion [6], and Average DCC [7].

Insertion evaluates the speed at which the classification score $\sigma(\cdot)$ increases by sequentially adding the pixels p and channels c with high importance in the importance map $\mathcal{I}$ to the image. In contrast, deletion evaluates the speed at which the classification score $\sigma(\cdot)$ decreases by sequentially removing the pixels p and channels c with low importance from the image. Both are quantified using the AUC (Area Under Curve) of the classification score, where higher AUC values for Insertion and lower AUC values for Deletion indicate better performance.

The average DCC is calculated by the harmonic mean of the three metrics: Coherency, Minimum Complexity, and Average drop, and takes a high value when the importance map accurately captures objects and is simple without including the background.

4.3 Experimental Data

The following two public datasets are used.

Table 1: Performance comparison on GTSRB dataset using Insertion, Deletion, and Average DCC. The best performance in each metric is indicated in bold.

Method	Insertion	Deletion	Average DCC
MC-RISE ($N^{\mathrm{mask}} = 1000$)	0.64	0.22	0.60
MC-RISE ($N^{\mathrm{mask}} = 5000$)	0.73	0.18	0.62
RaCM ($N^{\mathrm{mask}} = 1000$)	0.78	0.14	0.67
RaCM ($N^{\mathrm{mask}} = 5000$)	**0.83**	**0.13**	0.71
RaCM + Grad-CAM ($N^{\mathrm{mask}} = 1000$, $\theta = 30\%$)	0.64	0.22	**0.82**
RaCM + Grad-CAM ($N^{\mathrm{mask}} = 1000$, $\theta = 50\%$)	0.73	0.18	0.81
RaCM + Grad-CAM ($N^{\mathrm{mask}} = 5000$, $\theta = 30\%$)	0.67	0.21	**0.82**
RaCM + Grad-CAM ($N^{\mathrm{mask}} = 5000$, $\theta = 50\%$)	0.77	0.17	0.81

GTSRB [5] is an image dataset of 43 types of German traffic signs. The image size is resized to $N^{\mathrm{h}} \times N^{\mathrm{w}} = 96 \times 96$. Since traffic signs have characteristic colors and shapes, it is considered that specific pixels p and channels c contribute to the classification. The visualization target model is a VGG-16 model [11] trained with the same settings as MC-RISE [3], and all test data of the 43 classes are evaluated.

ImageNet [9] is a large-scale image dataset consisting of 1000 classes. The visualization target model is a pre-trained ResNet50 model [4], and the evaluation is conducted using test data of 15 classes (Acoustic guitar, Airliner, American chameleon, Banana, Goldfish, Bell paper, Cleaver, Cucumber, French horn, Lemon, Orange, Sports car, Strawberry, Red fox, Street sign) with distinctive colors.

4.4 Experimental results

The evaluation results for the GTSRB and ImageNet datasets are depicted in Tables 1 and 2. The Insertion, Deletion, and Average DCC values for each dataset are averaged over all classes respectively after calculating for each test image.

As shown in Table 1, in the GTSRB dataset, there is a significant performance difference in MC-RISE between a small number of masks ($N^{\mathrm{mask}} = 1000$) and a large number of masks ($N^{\mathrm{mask}} = 5000$), while the performance of RaCM does not change much. This indicates that RaCM can efficiently capture the target objects even with fewer masks. Additionally, RaCM+Grad-CAM significantly lowers the performance of the proposed RaCM in Insertion, suggesting that pre-refining the regions is limited more than necessary. Meanwhile, since the exploring regions become smaller, Complexity in Average DCC tends to be smaller, and the metric tends to be higher. In contrast, in the ImageNet dataset, although the performance is lower compared to GTSRB, the proposed RaCM significantly outperforms MC-RISE.

Fig. 5 depicts an example of visualizing the importance map for a STOP sign with red components from the GTSRB test data using 1000 masks. The

Table 2: Results of each method on ImageNet data

Method	Insertion	Deletion	Average DCC
MC-RISE ($N^{\text{mask}} = 1000$)	0.26	0.51	0.23
MC-RISE ($N^{\text{mask}} = 5000$)	0.28	0.53	0.22
RaCM ($N^{\text{mask}} = 1000$)	**0.45**	**0.26**	0.39
RaCM ($N^{\text{mask}} = 5000$)	0.44	0.28	**0.40**
RaCM + Grad-CAM ($N^{\text{mask}} = 1000$, $\theta = 30\%$)	0.34	0.47	0.22
RaCM + Grad-CAM ($N^{\text{mask}} = 1000$, $\theta = 50\%$)	0.40	0.40	0.32
RaCM + Grad-CAM ($N^{\text{mask}} = 5000$, $\theta = 30\%$)	0.37	0.48	0.23
RaCM + Grad-CAM ($N^{\text{mask}} = 5000$, $\theta = 50\%$)	0.44	0.41	0.33

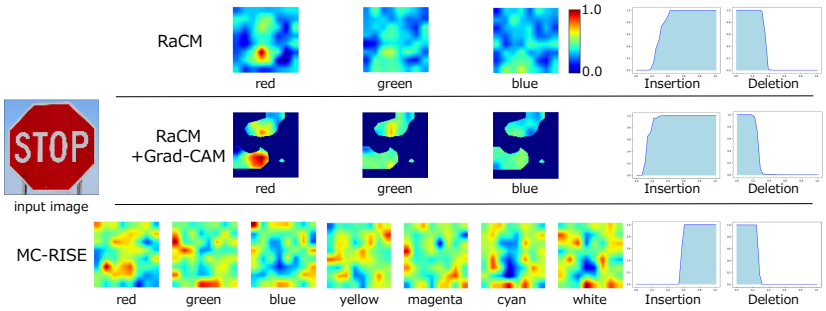

Fig. 5: Example of importance maps for the STOP sign in GTSRB (number of masks $N^{\text{mask}} = 1000$, $\theta = 30\%$)

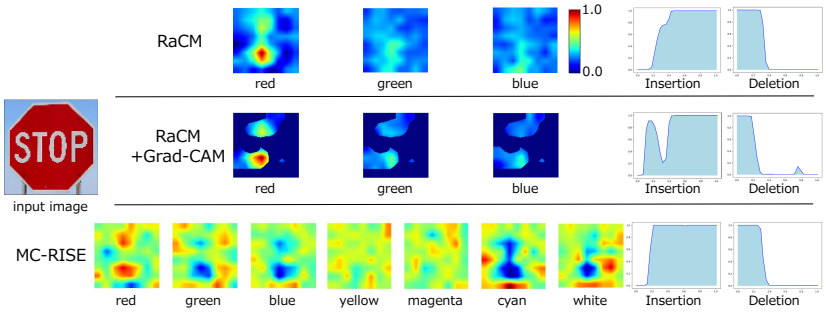

Fig. 6: Example of importance maps for the STOP sign in GTSRB (number of masks $N^{\text{mask}} = 5000$, $\theta = 30\%$)

results for MC-RISE visualize all the seven colors used. While MC-RISE captures the red component of the sign, it also captures unrelated parts of the sign with colors other than red. In contrast, the proposed RaCM strongly captures the red component and does not react strongly with other color components. Fig. 6 depcits the results for the same image using 5000 masks. Compared to 1000 masks, MC-RISE captures the red component more strongly, but still reacts

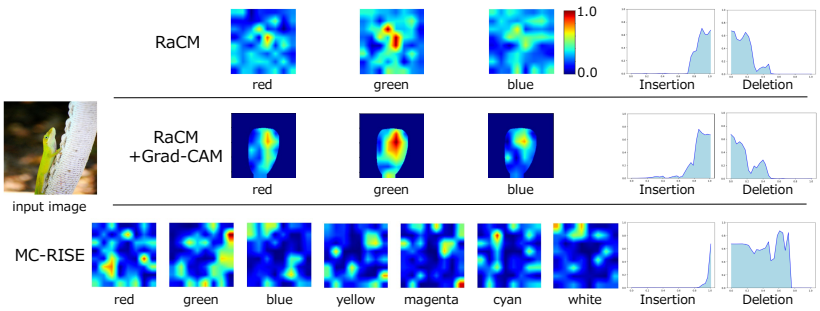

Fig. 7: Example of importance maps for the American chameleon in ImageNet (number of masks $N^{\text{mask}} = 1000$, $\theta = 30\%$)

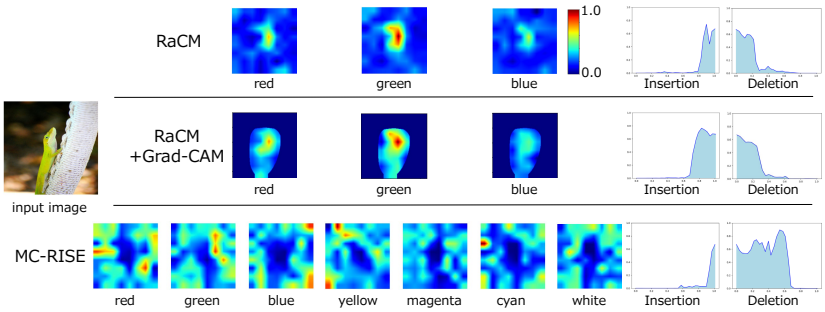

Fig. 8: Example of importance maps for the American chameleon in ImageNet (number of masks $N^{\text{mask}} = 5000$, $\theta = 30\%$)

strongly to other color components. In contrast, RaCM does not change significantly from 1000 masks, but with 5000 masks, the values of unrelated parts are lower, resulting in a sharper importance map.

Examples of the results for an American chameleon from ImageNet are depicted in Figs. 7 and 8. In MC-RISE, the areas where the object is located are not evaluated as green or yellow, but the surrounding background areas are evaluated. This suggests that MC-RISE has the problem discussed in Section 3.3, where if the target object is not masked, that mask is evaluated. In contrast, RaCM captures the target object's region and captures the important green component. These results indicate that MC-RISE can detect the target if it is large and centrally located in the image, as in GTSRB, but may have difficulty if the target object is small, as in ImageNet, where the background is likely to be evaluated. This explains the low evaluation values of MC-RISE in Table 2. In contrast, RaCM did not evaluate the mask if the target object was not masked and could find important parts with fewer masks.

Overall, these experimental results indicate that the proposed RaCM method could be an effective solution for pixel-channel-wise model interpretation, providing a combination of high performance and computational efficiency.

5 Conclusion

In this study, we proposed RaCM, a method to visualize the importance of pixels and color channels in images by randomly excluding color components and evaluating the significance of image-derived colors. RaCM generates random channel-pass masks that pass or exclude the color channels of each pixel, allowing for a more efficient and accurate evaluation of the importance of the original RGB values compared to the existing MC-RISE method. Through extensive experiments using GTSRB and ImageNet datasets, we demonstrated that RaCM outperforms existing methods in terms of both visualization performance and computational efficiency. These results suggest that RaCM is a promising approach for pixel-channel-wise model interpretation, providing a combination of high performance and efficiency in visualizing the importance of color channels in black-box models.

The experimental evaluation of proposed methods for HSV images or image sets, where it is difficult for humans to intuitively interpret which channels and locations influence the model's decisions, will be an important area of future work.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP23K11218.

A Derivation of approximated importance map for RaCM

The conditional probabilities $P[M^{\mathrm{CM}}|M^{\mathrm{CM}}(p,c)=1]$ and $P[M^{\mathrm{CM}}|M^{\mathrm{CM}_0}(p)=1]$ are derived as follows:

$$\begin{aligned} P[M^{\mathrm{CM}}|M^{\mathrm{CM}}(p,c)=1] &= \frac{P[M^{\mathrm{CM}}, M^{\mathrm{CM}}(p,c)=1]}{P[M^{\mathrm{CM}}(p,c)=1]} \\ &= \frac{P[M^{\mathrm{CM}}(p,c)=1|M^{\mathrm{CM}}]P[M^{\mathrm{CM}}]}{P[M^{\mathrm{CM}}(p,c)=1]} \\ &= \frac{M^{\mathrm{CM}}(p,c)P[M^{\mathrm{CM}}]}{P[M^{\mathrm{CM}}(p,c)=1]}, \end{aligned} \tag{13}$$

where $P[M^{\mathrm{CM}}(p,c)=1|M^{\mathrm{CM}}]=1$ when $M^{\mathrm{CM}}(p,c)=1$ and 0 otherwise.

$$\begin{aligned}P[M^{\mathrm{CM}}|M^{\mathrm{CM}_0}(p)=1] &= \frac{P[M^{\mathrm{CM}}, M^{\mathrm{CM}_0}(p)=1]}{P[M^{\mathrm{CM}_0}(p)=1]} \\ &= \frac{P[M^{\mathrm{CM}_0}(p)=1|M^{\mathrm{CM}}]P[M^{\mathrm{CM}}]}{P[M^{\mathrm{CM}_0}(p)=1]} \\ &= \frac{M^{\mathrm{CM}_0}(p)P[M^{\mathrm{CM}}]}{P[M^{\mathrm{CM}_0}(p)=1]}. \end{aligned} \tag{14}$$

If each channel c is randomly selected independently with a probability of $\frac{1}{2}$, the number of combinations where at least one of the N^{ch} channels is selected is $2^{N^{\mathrm{ch}}}-1$. Furthermore, of the $2^{N^{\mathrm{ch}}}$ possible combinations, the number of combinations in which a specific channel c is selected is $\frac{2^{N^{\mathrm{ch}}}}{2}=2^{N^{\mathrm{ch}}-1}$. Therefore, assuming that the probability of selecting a pixel p is P^{pix}, the probability that a pixel p and channel c are selected to be passed is as follows:

$$P[M^{\mathrm{CM}}(p,c)=1]=\frac{2^{N^{\mathrm{ch}}-1}\times P^{\mathrm{pix}}}{2^{N^{\mathrm{ch}}}-1}. \tag{15}$$

For example, in the case of RGB channels, i.e., $N^{\mathrm{ch}}=3$, there are seven types of color: red $\boldsymbol{c}_1=(1,0,0)$, green $\boldsymbol{c}_2=(0,1,0)$, blue $\boldsymbol{c}_3=(0,0,1)$, yellow $\boldsymbol{c}_4=(1,1,0)$, magenta $\boldsymbol{c}_5=(1,0,1)$, cyan $\boldsymbol{c}_6=(0,1,1)$, and white $\boldsymbol{c}_7=(1,1,1)$. Then, the probability of selecting a pixel p and a channel c is $P[M^{\mathrm{CM}}(p,c)=1]=\frac{4P^{\mathrm{pix}}}{7}$ and the probability of not selecting a pixel is $P[M^{\mathrm{CM}_0}(p)=1]=1-P^{\mathrm{pix}}$.

Using Eqs. 13, 14, and 15, we can expand Eq. 9 and estimate it using Monte Carlo samples of masks as follows:

$$\begin{aligned}&\mathcal{I}_k^{\mathrm{CM}}(p,c) \\ &= \int \left(P[M^{\mathrm{CM}}|M^{\mathrm{CM}}(p,c)=1]-P[M^{\mathrm{CM}}|M^{\mathrm{CM}_0}(p)=1]\right)\sigma^k\left(I^{\mathrm{CM}}(p;M^{\mathrm{CM}})\right)dM^{\mathrm{CM}} \\ &= \int \left(\frac{M^{\mathrm{CM}}(p,c)}{P[M^{\mathrm{CM}}(p,c)=1]}-\frac{M^{\mathrm{CM}_0}(p)}{P[M^{\mathrm{CM}_0}(p)=1]}\right)P[M^{\mathrm{CM}}]\sigma^k\left(I^{\mathrm{CM}}(p;M^{\mathrm{CM}})\right)dM^{\mathrm{CM}} \\ &= \mathbb{E}_{M^{\mathrm{CM}}}\left[\left(\frac{2^{N^{\mathrm{ch}}-1}\times M^{\mathrm{CM}}(p,c)}{(2^{N^{\mathrm{ch}}}-1)P^{\mathrm{pix}}}-\frac{M^{\mathrm{CM}_0}(p)}{1-P^{\mathrm{pix}}}\right)\sigma^k\left(I^{\mathrm{CM}}(p;M^{\mathrm{CM}})\right)\right] \\ &\approx \frac{1}{N^{\mathrm{mask}}}\sum_{m=1}\left(\frac{2^{N^{\mathrm{ch}}-1}\times M_m^{\mathrm{CM}}(p,c)}{(2^{N^{\mathrm{ch}}}-1)P^{\mathrm{pix}}}-\frac{M_m^{\mathrm{CM}_0}(p)}{1-P^{\mathrm{pix}}}\right)\sigma^k\left(I^{\mathrm{CM}}(p;M_m^{\mathrm{CM}})\right) \end{aligned} \tag{16}$$

References

1. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (2018)

2. Desai, S., Ramaswamy, H.G.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 972–980 (2020)
3. Hatakeyama, Y., Sakuma, H., Konishi, Y., Suenaga, K.: Visualizing color-wise saliency of black-box image classification models. In: Asian Conference on Computer Vision (ACCV). pp. 189–205 (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
5. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: International Joint Conference on Neural Networks. No. 1288 (2013)
6. Petsiuk, V., Das, A., Saenko, K.: Rise: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (BMVC). pp. 1–13 (2018)
7. Poppi, S., Cornia, M., Cucchiara, L.B.R.: Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–6 (2021)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data. pp. 1135–1144 (2016)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015)
10. 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: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
12. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
13. Zhang, Q., Rao, L., Yang, Y.: Group-cam: Group score-weighted visual explanations for deep convolutional networks. arXiv preprint arXiv:2103.13859 pp. 1–9 (2021)