Supplementary Material for "Visualizing Color-wise Saliency of Black-Box Image Classification Models"

Yuhki Hatakeyama[1*], Hiroki Sakuma[1], Yoshinori Konishi[1], and Kohei Suenaga[2]

SenseTime Japan, 4F, Oike Koto Building, 324 Oikeno-cho, Nakagyo-ku, Kyoto, Japan
{hatakeyama,sakuma,konishi}@sensetime.jp
Graduate School of Informatics, Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
ksuenaga@gmail.com

A Derivation of Equation (4)

We rewrite the definition of $S^{\mathrm{PN}}_{i,l}(\lambda)$ (Equation (3)) as the sum over all possible masks m;

$$\begin{aligned} S^{\mathrm{PN}}_{i,l}(\lambda) &= \mathbb{E}_m[M(i \odot m, l) | m(\lambda) = 1] - \mathbb{E}_m[M(i \odot m, l) | m(\lambda) = 0] \\ &= \sum_m M(I \odot m, l) \\ &\quad \times (P[X = m | X(\lambda) = 1] - P[X = m | X(\lambda) = 0]) \ . \end{aligned} \tag{A.1}$$

$P[X = m|X(\lambda) = 1]$ and $P[X = m|X(\lambda) = 0]$ are expressed as

$$\begin{aligned} P[X = m | X(\lambda) = 1] &= \frac{P[X = m, X(\lambda) = 1]}{P[X(\lambda) = 1]} \\ &= \begin{cases} 0 & \text{if } m(\lambda) = 0 \\ \frac{P[X=m]}{P[X(\lambda)=1]} & \text{if } m(\lambda) = 1 \end{cases} \\ &= \frac{m(\lambda) P[X = m]}{p} \ , \end{aligned} \tag{A.2}$$

$$\begin{aligned} P[X = m | X(\lambda) = 0] &= \begin{cases} \frac{P[X=m]}{P[X(\lambda)=0]} & \text{if } m(\lambda) = 0 \\ 0 & \text{if } m(\lambda) = 1 \end{cases} \\ &= \frac{(1 - m(\lambda)) P[X = m]}{1 - p} \ , \end{aligned} \tag{A.3}$$

where $p = P[X(\lambda) = 1] = 1 - P[X(\lambda) = 0]$. Substituting these expressions in the expression of $S^{\mathrm{PN}}_{i,l}(\lambda)$ (Equaiton (A.1)), we obtain Equation (4);

$$\begin{aligned} S^{\mathrm{PN}}_{i,l}(\lambda) &= \sum_m \left(\frac{m(\lambda)}{p} - \frac{1 - m(\lambda)}{1-p} \right) M(I \odot m, l) P[X = m] \\ &= \sum_m \frac{m(\lambda) - p}{p(1-p)} M(i \odot m, l) P[X = m] \ . \end{aligned} \tag{A.4}$$

B Derivation of Equation (9)

Following the similar derivation in §A, the definition of $S^{\mathrm{MC}}_{i,l}(\lambda)$ (Equation (8)) can be expressed as

$$\begin{aligned} S^{\mathrm{MC}}_{i,l}(\lambda, k) &:= \mathbb{E}_{m^{\mathrm{c}} \sim \mathcal{M}_c}[M(i'(\lambda; m^{\mathrm{c}}))|m^{\mathrm{c}}(\lambda, k) = 1] \\ &\quad - \mathbb{E}_{m^{\mathrm{c}} \sim \mathcal{M}_c}[M(i'(\lambda; m^{\mathrm{c}}))|m^{(0)}(\lambda) = 1] \\ &= \sum_m M(i'(\lambda; m^{\mathrm{c}}), l) \\ &\quad \times \Big(P[X = m^{\mathrm{c}}|X(\lambda, k) = 1] - P[X = m^{\mathrm{c}}|X^{(0)}(\lambda) = 1] \Big) \\ &= \sum_m \left(\frac{m^{\mathrm{c}}(\lambda, k)}{P[X(\lambda, k) = 1]} - \frac{m^{(0)}(\lambda)}{P[X^{(0)}(\lambda) = 1]]} \right) \\ &\quad \times M(i'(\lambda; m^{\mathrm{c}}), l) P[X = m] \ . \end{aligned} \tag{A.5}$$

In the mask generation process, whether a pixel is masked or retained is randomly determined with the masking probability p_{mask}, and for a masked pixel, the masking color is sampled from the uniform distribution over K colors; hence, $P[X(\lambda, k) = 1] = p_{\mathrm{mask}} \frac{1}{K}$ and $P[X^{(0)}(\lambda) = 1] = 1 - p_{\mathrm{mask}}$. Substituting them in the expression of $S^{\mathrm{MC}}_{i,l}(\lambda)$ (Equation (A.5)) yields Equation (9);

$$S^{\mathrm{MC}}_{i,l}(\lambda, k) = \mathbb{E}_{m^{\mathrm{c}} \sim \mathcal{M}_c} \left[\left(\frac{m^{\mathrm{c}}(\lambda, k)}{p_{\mathrm{mask}}/K} - \frac{m^{(0)}(\lambda)}{1 - p_{\mathrm{mask}}} \right) M(i'(\lambda; m^{\mathrm{c}})) \right] \ . \tag{A.6}$$

C Proof of Proposition 1

Let $\lambda \in \Lambda$ be a fixed pixel and A be the set of all possible masks. We define a disjoint partition of A by $A^+ = \{m \in A | m(\lambda) = 1\}$ and $A^- = \{m \in A | m(\lambda) = 0\}$. From the definition of $S^{\mathrm{PN}}_{i,l}(\lambda)$ (Equation (3)), we get

$$\begin{aligned} S^{\mathrm{PN}}_{i,l}(\lambda) = &\sum_{m \in A^+} M(I \odot m, l) P[X = m | X(\lambda) = 1] \\ &- \sum_{m \in A^-} M(I \odot m, l) P[X = m | X(\lambda) = 0] \end{aligned} \tag{A.7}$$

because $P[X = m|X(\lambda) = 1] = 0$ if $m(\lambda) = 0$ and vice versa.

Let $F_\lambda : A \to A$ be the function which flips the mask value at pixel λ. F_λ induces a one-to-one correspondence between the masks in A^+ and A^-. Therefore, Equation (A.7) is expressed as

$$\begin{aligned} S^{\mathrm{PN}}_{i,l}(\lambda) = &\sum_{m \in A^+} M(I \odot m, l) P[X = m|X(\lambda) = 1] \\ &- \sum_{m \in A^+} M(I \odot F_\lambda(m), l) P[X = F_\lambda(m)|X(\lambda) = 0] \ . \end{aligned} \tag{A.8}$$

We can rewrite $P[X = F_\lambda(m)|X(\lambda) = 0]$ as

$$\begin{aligned} P[X = F_\lambda(m)|X(\lambda) = 0] &= \delta_{F_\lambda(m)(\lambda),0} \prod_{\kappa \in \Lambda \setminus \{\lambda\}} p^{m(\kappa)} (1-p)^{1-m(\kappa)} \\ &= \delta_{m(\lambda),1} \prod_{\kappa \in \Lambda \setminus \{\lambda\}} p^{m(\kappa)} (1-p)^{1-m(\kappa)} \\ &= P[X = m|X(\lambda) = 1] \ , \end{aligned} \tag{A.9}$$

where $\delta_{i,j}$ is the Kronecker delta and p is the masking probability for a pixel. Therefore, Equation (A.8) is rewritten as

$$\begin{aligned} S^{\mathrm{PN}}_{i,l}(\lambda) = &\sum_{m \in A^+} \{M(I \odot m, l) - M(I \odot F_\lambda(m), l)\} \\ &\times P[X = m|X(\lambda) = 1] \ . \end{aligned} \tag{A.10}$$

Since the premise in Proposition 1 is expressed as

$$M(I \odot m, l) = M(I \odot F_\lambda(m), l) \text{ for all } m \in A^+ \ , \tag{A.11}$$

we obtain $S^{\mathrm{PN}}_{i,l}(\lambda) = 0$.

D Additional Experiments with Person ReID model

This section presents the application of MC-RISE to a metric-learning-based person re-identification (ReID) model.

A metric-learning-based person ReID model takes an image i of a person as input and outputs the feature vector v_i for the input image. The model is trained so that, if it is given a pair of images i and i', then the distance between v_i and $v_{i'}$ is small if i and i' are likely to be the images of the same person. We designate a set of gallery images; at inference time, we compare the feature vector of a query image with those of the gallery images and retrieve a gallery image that belongs to the same person as the query.

We adapted MC-RISE as follows to apply it to a person ReID model.

- Unlike a standard classification task where an image belongs to a unique class, a person ReID task has multiple correct gallery images for one query image in general. In this experiment, we only consider the gallery image of the top-1 match to a query image i as the correct label; the feature distance to the top-1 match image is used as the output of the black-box model $M(i, l)$, that is used by MC-RISE. Hence, MC-RISE visualizes how the *similarity* to the top-1 match image responds to the color masking of an input image.
- As for the definition of feature distance, the simple Euclidean distance between feature vectors is not appropriate because the weighted sum of color masks would be dominated by outlier samples with large feature distance, resulting in the uninterpretable saliency maps. We computed $M(i, l)$ by the following formula:

$$M(i,l) = \exp\left(-\frac{d(f_i, f_{\text{match}})}{d_0}\right). \tag{A.12}$$

  Here, $d(f_i, f_{\text{match}})$ is the feature distance between an input image i and the top-1 match image, and d_0 is a typical scale of the distance, for which we used the feature distance between the original query image and the top-1 match image. The resulting saliency maps are not much affected by the outliers since Equation (A.12) becomes nearly zero for an outlier with a large distance.

In our experiment, we visualized the color saliency maps for the Market-1501 dataset [1]. For the evaluation, we used the pretrained OSNet($\times$1.0) [2] model provided by Torchreid library [3]. The parameters for MC-RISE were the same as §5.1 except that the masking probability was set to 0.1.

Fig. 1 shows the saliency maps generated for the Market-1501 dataset. For most of the query images, the saliency maps have negative values on the entire body as the top row sample shows. This suggests that the model compares the whole parts of the body in a query image with that of the gallery images; if the colors of corresponding parts disagree, it largely diminishes the similarity between images. However, for some queries such as the ones in the middle or the bottom row, the saliency maps indicate that the model pays close attention to a specific part of the body (e.g., head in the middle row sample) or the specific color of clothing (e.g., green color clothing in the bottom row sample). These results demonstrate that MC-RISE can also be applied to metric learning problems, such as person ReID task, and can visualize the characteristics of the model's decision.

E MC-RISE with $K = 8$

In the experiments with GTSRB dataset in §5, we applied MC-RISE by setting the number of colors to 5 (i.e., $K = 5$). Fig. 2 shows several saliency maps generated by MC-RISE wherein we set K to 8; the other settings are kept the same as in §5.

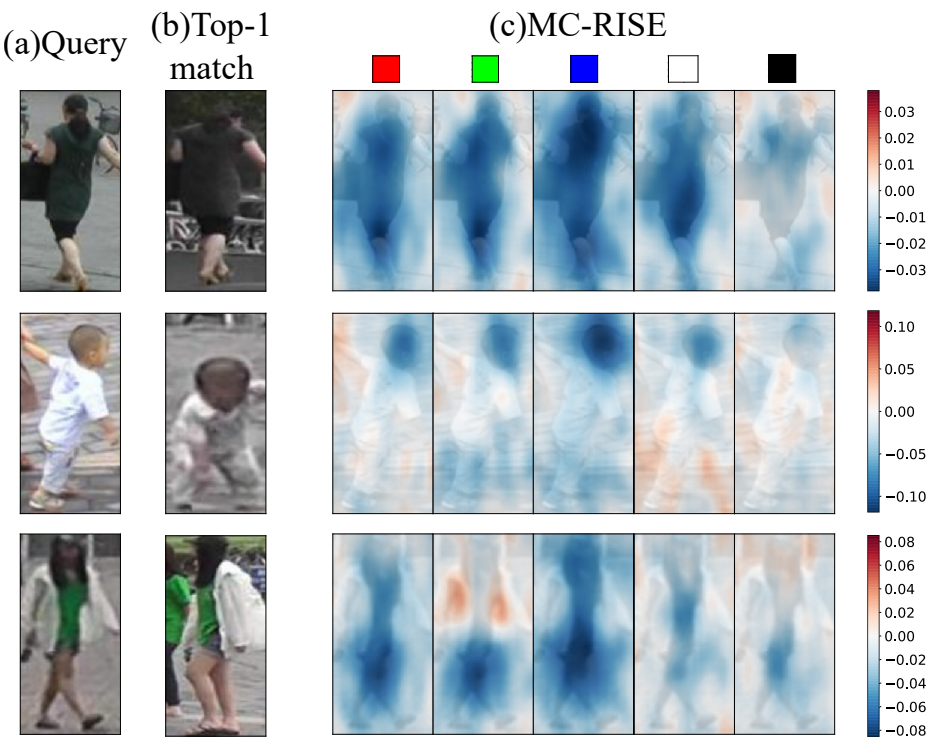

Fig. 1. The color saliency maps generated by MC-RISE for the Market-1501 dataset. The saliency maps in (c) visualize the color-wise saliency for the similarity between (a) a query image and (b) the top-1 match image in gallery images. All queries are correctly matched by the model. Best viewed in color.

Although the tendency of the saliency maps is by and large the same as that in §5, it is worth noting that, by using more colors, we can read out more information from the saliency map in the bottom row in Fig. 2 than that in Fig. 4 in §5. We can observe that, in addition to red at the center of the sign, stronger yellow and magenta at the center would make the confidence more solid.

F Pseudocode of MC-RISE

The pseudocode of MC-RISE is presented in Algorithm 1.

References

1. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)

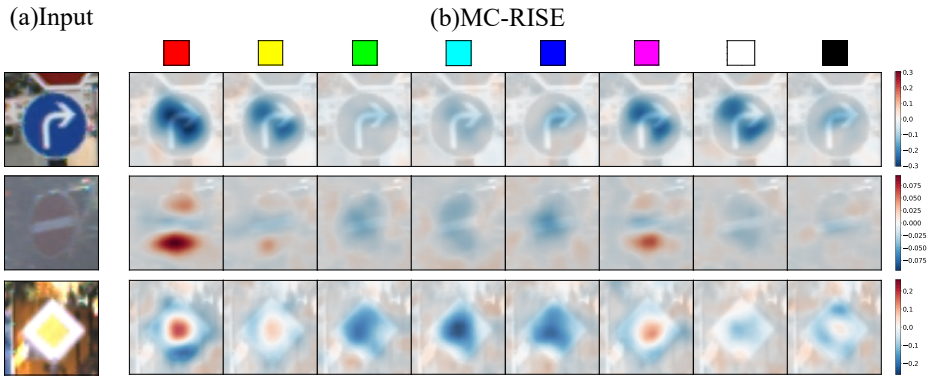

Fig. 2. Saliency maps generated by MC-RISE with $K = 8$ for GTSRB dataset.

2. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
3. Zhou, K., Xiang, T.: Torchreid: A library for deep learning person re-identification in pytorch. arXiv preprint arXiv:1910.10093 (2019)

Algorithm 1: MC-RISE algorithm

Input: Input image $i(\lambda)$, Target label l, Black-box model M, Color set $\{c_k\}_{k=1}^{K}$, The number of masks N, Low-resolution mask size $h \times w$

Output: Color saliency maps $S_{i,l}^{\mathrm{MC}}(\lambda, k)$

$S^{\mathrm{raw}}(\lambda, k) \leftarrow 0$, $S^{\mathrm{baseline}}(\lambda) \leftarrow 0$ for all λ, k

for $n = 1$ *to* N **do**

  `// generating color masks`

  $m_{\mathrm{mask}} \leftarrow$ randomly sample $h \times w$ binary mask with the masking probability p_{mask}.

  for *pixel* λ *in* $h \times w$ *image* **do**

    `// eq.(6)`

    $m_{\mathrm{low}}(\lambda, k) \leftarrow 0$ for all $k = 1 \dots K$

    if $m_{\mathrm{mask}}(\lambda) = 1$ **then**

      $k' \leftarrow$ randomly sample the index of masking color from $\{1 \dots K\}$

      $m_{\mathrm{low}}(\lambda, k') \leftarrow 1$

  $m_n^c \leftarrow \mathit{bilinear\_interpolation}(m_{\mathrm{low}})$

  $m_n^c \leftarrow \mathit{random\_shift}(m_n^c)$

  $m_n^{(0)}(\lambda) \leftarrow 1 - \sum_{k=1}^{K} m_n^c(\lambda, k)$ for all λ `// eq.(7)`

  `// computing saliency maps; eq.(10)`

  $i'(\lambda) \leftarrow i(\lambda) m_n^{(0)}(\lambda) + \sum_{k=1}^{K} c_k m_n^c(\lambda, k)$ for all λ `// eq.(8)`

  $p_{\mathrm{out}} \leftarrow M(i', l)$

  $S^{\mathrm{raw}}(\lambda, k) \mathrel{+}= \frac{K m_n^c(\lambda, k)}{p_{\mathrm{mask}}} p_{\mathrm{out}}$ for all λ, k

  $S^{\mathrm{baseline}}(\lambda) \mathrel{+}= \frac{m_n^{(0)}(\lambda)}{(1 - p_{\mathrm{mask}})} p_{\mathrm{out}}$ for all λ

$S_{i,l}^{\mathrm{MC}}(\lambda, k) \leftarrow (S^{\mathrm{raw}}(\lambda, k) - S^{\mathrm{baseline}}(\lambda))/N$ for all λ, k