BOREx: Bayesian-Optimization–Based Refinement of Saliency Map for Image- and Video-Classification Models

Supplementary Material

Atsushi Kikuchi, Kotaro Uchida, Masaki Waga[0000−0001−9360−7490], and Kohei Suenaga[0000−0002−7466−8789]

Kyoto University

A The detail of experimental environment

We implemented Algorithm 2 using Python 3.6.12 and PyTorch 1.6.0. The experiments on image classification are conducted on a GPU workstation with 3.60 GHz Intel Core i7-6850K, 12 CPUs, NVIDIA Quadro P6000, and 32GB RAM that runs Ubuntu 20.04.2 LTS (64 bit) and CUDA 11.0. The experiments on video classification are conducted on a GPU workstation with 3.00 GHz Intel Xeon E5-2623 v3, 16 CPUs, NVIDIA Tesla P100, and 500GB RAM that runs Ubuntu 16.04.3 LTS (64 bit). In the implementation of Algorithm 2, we used Matérn kernel defined as follows:

$$k((\lambda_1, r_1), (\lambda_2, r_2)) := \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} d'}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} d'}{l} \right),$$

where d' is the Euclidean distance between (λ_1, r_1) and (λ_2, r_2) (i.e., $\sqrt{d(\lambda_1, \lambda_2)^2 + (r_1 - r_2)^2}$ where $d(\lambda_1, \lambda_2)$ is the Euclidean distance between λ_1 and λ_2 in the input image); ν and l are positive parameters that control the shape of the function; Γ is the gamma function; and K_ν is a modified Bessel function [Abramowitz and Stegun(1972)]. Mokuwe et al. [Mokuwe et al.(2020)] used $\nu = 2.5$ and $l = 12$ in their implementation. We use the Matérn kernel with $\nu = 1.5$ and $l = 12$.

B Results of additional experiments

B.1 Effect of increasing N in Algorithm 2

We executed Algorithm 2 with different N, the number of iterations for Gaussian process regression. With more iterations, the estimated μ is expected to be more precise. However, the time spent in one iteration in the later iterations tends to be longer because there are more observations to fit.

Tables 1 and 2 show the result. We observe the significant improvement in the insertion metric only when we increased N from 10 to the other values; in

Table 1: Comparison of the results produced by different number of N in Algorithm 2. Each result uses an input saliency map generated by RISE with 100 masks. "BOn" represents Algorithm 2 with $N = n$. Each column represents the following: "Metric" for the metric; "Base." for the baseline; "Comp." for the compared method; "Stat." for the test statistic of Wilcoxon test; "p-val." for the p-value. One asterisk indicates $p < 0.05$; two asterisks indicates $p < 0.001$.

Metric	Base.	Comp.	Stat.	p-val.
Insertion	BO10	BO50	71390	**7.042e-36****
	BO10	BO80	72357	**4.023e-38****
	BO10	BO100	72463	**2.261e-38****
	BO50	BO80	45042	8.156e-02
	BO50	BO100	45142	7.542e-02
	BO80	BO100	42923	3.066e-01
Deletion	BO10	BO50	13676	**2.937e-32****
	BO10	BO80	10935	**1.838e-38****
	BO10	BO100	9708	**2.007e-41****
	BO50	BO80	35484	**4.454e-03***
	BO50	BO100	32929	**1.132e-04****
	BO80	BO100	37466	3.721e-02
F-measure	BO10	BO50	69660	**4.825e-32****
	BO10	BO80	70810	**1.442e-34****
	BO10	BO100	70879	**1.010e-34****
	BO50	BO80	48741	**1.606e-03***
	BO50	BO100	49295	**7.387e-04****
	BO80	BO100	41233	5.806e-01

the other cases, improvement nor degradation is not concluded (Table 2). For the other metrics, increasing N from 50 to the other values is also concluded to be effective. In any metrics, increasing N from 80 to 100 was not concluded to be effective. This result suggests that increasing N, which incurs time for executing Algorithm 1, is effective; however, the merit of increasing the number beyond certain number (here 50) is limited.

B.2 Effect of the quality of an input saliency map on the output of Algorithm 2

We executed Algorithm 2 with input maps generated by RISE with different numbers of masks. This experiment is to study the effect of the quality of an input saliency map on the output because the quality of an input saliency map is expected to be higher with more masks.

Table 3 shows the result. Significant improvement is observed (1) in the insertion metric when we increased the number of masks from 0 to 300 or more and (2) in the F-measure when we increased the number of masks from 0 to 300. We cannot conclude the significant improvement in the other cases. The two-sided test we conducted, whose result is presented in Table 4, does not

Table 2: Two-sided test to compare the results produced by different number of N in Algorithm 2. Each result uses an input saliency map generated by RISE with 100 masks.

Metric	Base.	Comp.	Stat.	p-val.
Insertion	BO10	BO50	12046	**1.408e-35****
	BO10	BO80	11079	**8.046e-38****
	BO10	BO100	10973	**4.522e-38****
	BO50	BO80	38394	1.631e-01
	BO50	BO100	38294	1.508e-01
	BO80	BO100	40513	6.131e-01
Deletion	BO10	BO50	13676	**5.875e-32****
	BO10	BO80	10935	**3.677e-38****
	BO10	BO100	9708	**4.013e-41****
	BO50	BO80	35484	**8.907e-03***
	BO50	BO100	32929	**2.264e-04****
	BO80	BO100	37466	7.442e-02
F-measure	BO10	BO50	13776	**9.650e-32****
	BO10	BO80	12626	**2.884e-34****
	BO10	BO100	12557	**2.020e-34****
	BO50	BO80	34695	**3.212e-03***
	BO50	BO100	34141	**1.477e-03***
	BO80	BO100	41233	8.388e-01

conclude that the quality of the saliency maps measured in these metrics differ among the input saliency maps generated by RISE with different numbers of masks, which implies that the quality is not degraded by increasing the number of masks. These results back our conclusion of BOREx is effective to improve a low-quality saliency map in terms of the insertion metric.

C Examples of saliency maps

C.1 Examples in which BOREx successfully improves input images

Figures 1–4 present examples of saliency maps generated by BOREx. We picked several examples in which BOREx successfully refines the quality of input images measured in the quantitative metrics. Compared to the input images generated by RISE, the saliency maps generated by BOREx localize the important regions better and less noisy.

C.2 Examples in which BOREx degraded the quality of input images

Figure 5 presents the examples in which BOREx degraded the input images measured in the quantitative metrics. We add explanations for each example.

The first row in Figure 5, which BOREx degraded the insertion metric, presents an example in which RISE successfully identifies the aeroplane, whereas BOREx wrongly identifies the ground in addition to the aeroplane. This is caused by the saliency map produced by RISE used as the prior; in the prior, the saliency of the ground in the image is high, which misled 2.

The saliency maps in the second row of Figure 5 are generated by the label "chair". BOREx identifies one of the chairs in the image, but not the other chair, degrading the insertion metric because identifying two chairs are needed to recognize a park bench. This is due to the issue of the limited shape of the masks used by BOREx discussed in Section 4.1. RISE looks successfully identifies both chairs.

BOREx degraded the F-measure metric for the image in the third row of Figure 5. The important region identified by BOREx concentrates around the lid of the bottle, whereas the region identified by RISE exists also on the body of the bottle. The PascalVOC dataset specifies the entire bottle as the correct answer, which leads to the poor value in the F-measure metric for the saliency map generated by BOREx. Deciding from the insertion and the deletion metrics, we guess that the model indeed considers the lid part as the salient region.

D Comparison with Grad-CAM++

Although the statistical tests in Section 4 do not conclude the effectiveness of BOREx measured in the insertion and the F-measure metrics to refine the saliency map produced by Grad-CAM++, there are some instances that indeed benefit from BOREx. Figure 6 presents several examples of the saliency maps in which BOREx outperforms Grad-CAM++.

Interestingly, if we let BOREx and Grad-CAM++ produce saliency maps for the same image with different labels, the saliency maps produced by Grad-CAM++ are often less sensitive to the change in the label than BOREx. For example, the first (resp., the second) row in Figure 6 are the saliency maps produced by BOREx and Grad-CAM++ with label "sofa" (resp., "chair"). The saliency maps produced by BOREx correctly identify the region that is important for classifying the image to the given label; however, the saliency maps produced by Grad-CAM++ is less focused to the given label than BOREx.

E Examples of saliency maps for video classifiers

Figures 7–9 present examples of saliency maps for a video classifier produced by BOREx and a naive extension of RISE. Compared with the saliency maps produced by RISE, the saliency maps generated by BOREx better localize the important parts. It is also observed that the salinecy maps produced by BOREx are comparable to those produced by Grad-CAM++ in spite of the black-box nature of BOREx. In Figure 8, we can observe that BOREx follows the skier better than Grad-CAM++.

References

Abramowitz and Stegun(1972). Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Government Printing Office, Washington, DC, USA, tenth printing edition, 1972. 1

Mokuwe et al.(2020). Mamuku Mokuwe, Michael Burke, and Anna Sergeevna Bosman. Black-box saliency map generation using bayesian optimisation. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE, 2020. 1

Table 3: Comparison of the results of Algorithm 2 with input saliency maps produced by RISE with different number of masks. N is set to 50 in each execution of Algorithm 2. "RISEn" represents input salinecy maps produced by RISE with n masks are used. Each column represents the following: "Metric" for the metric; "Base." for the baseline; "Comp." for the compared method; "Stat." for the test statistic of Wilcoxon test; "p-val." for the p-value. One asterisk indicates $p < 0.05$; two asterisks indicates $p < 0.001$.

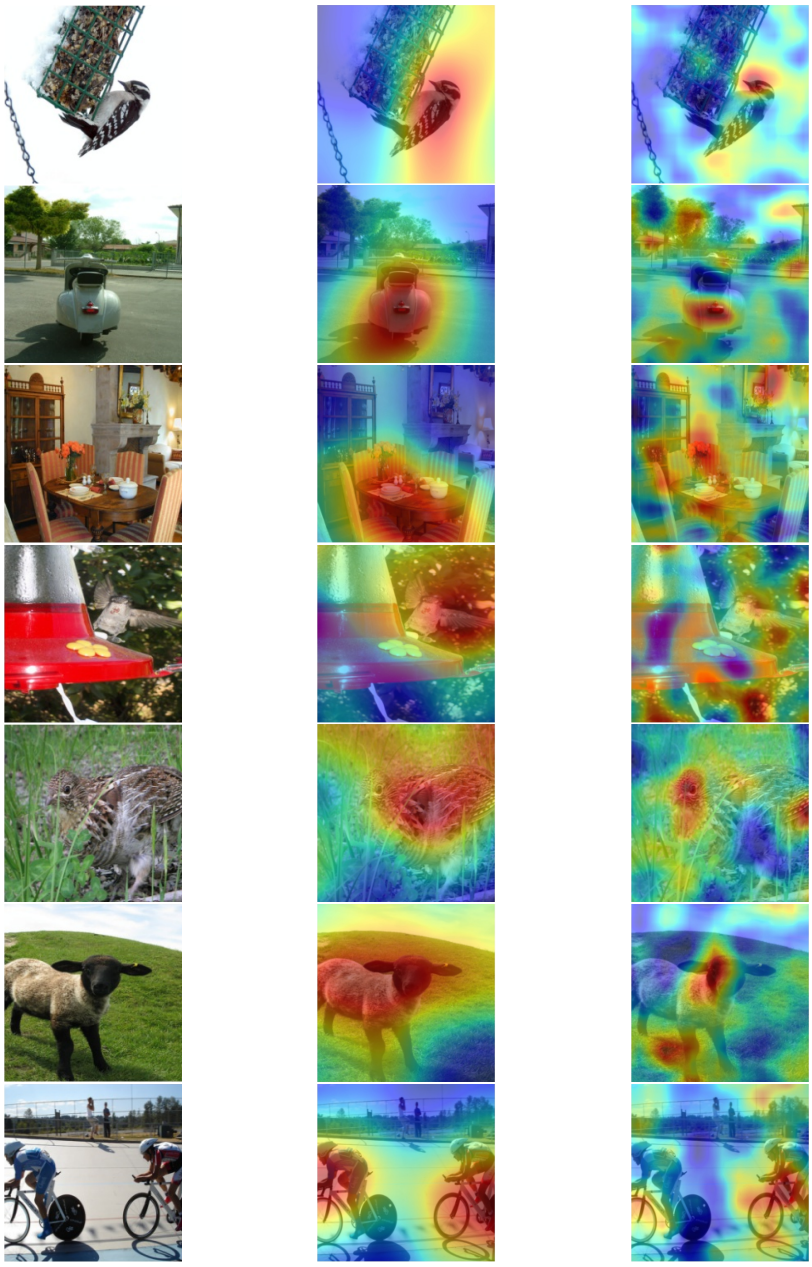
Metric	Base.	Comp.	Stat.	p-val.
Deletion	RISE0	RISE100	29768	8.453e-01
	RISE0	RISE300	19813	6.630e-01
	RISE0	RISE500	19660	6.202e-01
	RISE0	RISE1000	18734	3.491e-01
	RISE100	RISE300	18905	3.976e-01
	RISE100	RISE500	18656	3.277e-01
	RISE100	RISE1000	18016	1.773e-01
	RISE300	RISE500	19598	4.381e-01
	RISE300	RISE1000	19010	2.786e-01
	RISE500	RISE1000	18704	2.085e-01
F-meas.	RISE0	RISE100	30117	1.123e-01
	RISE0	RISE300	22294	**1.131e-02***
	RISE0	RISE500	20464	1.81777e-01
	RISE0	RISE1000	20559	1.636e-01
	RISE100	RISE300	19529	4.176e-01
	RISE100	RISE500	18431	7.307e-01
	RISE100	RISE1000	19061	5.568e-01
	RISE300	RISE500	18885	7.514e-01
	RISE300	RISE1000	17600	9.475e-01
	RISE500	RISE1000	20302	3.592e-01
Insertion	RISE0	RISE100	30962	4.525e-02
	RISE0	RISE300	22842	**3.566e-03***
	RISE0	RISE500	22848	**3.518e-03***
	RISE0	RISE1000	23337	**1.101e-03***
	RISE100	RISE300	20070	2.698e-01
	RISE100	RISE500	20650	1.473e-01
	RISE100	RISE1000	21153	7.709e-02
	RISE300	RISE500	20490	3.091e-01
	RISE300	RISE1000	21228	1.492e-01
	RISE500	RISE1000	20886	2.151e-01

Table 4: Two-sided test to compare Algorithm 2 with input saliency maps produced by RISE with different number of masks.

Metric	Base.	Comp.	Stat.	p-val.
Deletion	RISE0	RISE100	26177	3.09308e-01
	RISE0	RISE300	18690	6.73922e-01
	RISE0	RISE500	18843	7.59515e-01
	RISE0	RISE1000	18734	6.98164e-01
	RISE100	RISE300	18905	7.95128e-01
	RISE100	RISE500	18656	6.55416e-01
	RISE100	RISE1000	18016	3.54525e-01
	RISE300	RISE500	19598	8.76144e-01
	RISE300	RISE1000	19010	5.57116e-01
	RISE500	RISE1000	18704	4.17039e-01
F-meas.	RISE0	RISE100	25828	2.24636e-01
	RISE0	RISE300	16209	2.2610e-02*
	RISE0	RISE500	18039	3.63555e-01
	RISE0	RISE1000	17944	3.27184e-01
	RISE100	RISE300	18974	8.35268e-01
	RISE100	RISE500	18431	5.38648e-01
	RISE100	RISE1000	19061	8.86484e-01
	RISE300	RISE500	18885	4.97256e-01
	RISE300	RISE1000	17600	1.04953e-01
	RISE500	RISE1000	19319	7.18478e-01
Insertion	RISE0	RISE100	30962	4.5251e-02*
	RISE0	RISE300	22842	3.566e-03*
	RISE0	RISE500	22848	3.518e-03*
	RISE0	RISE1000	23337	1.101e-03*
	RISE100	RISE300	20070	2.698e-01
	RISE100	RISE500	20650	1.473e-01
	RISE100	RISE1000	21153	7.709e-02
	RISE300	RISE500	20490	3.091e-01
	RISE300	RISE1000	21228	1.492e-01
	RISE500	RISE1000	20886	2.151e-01

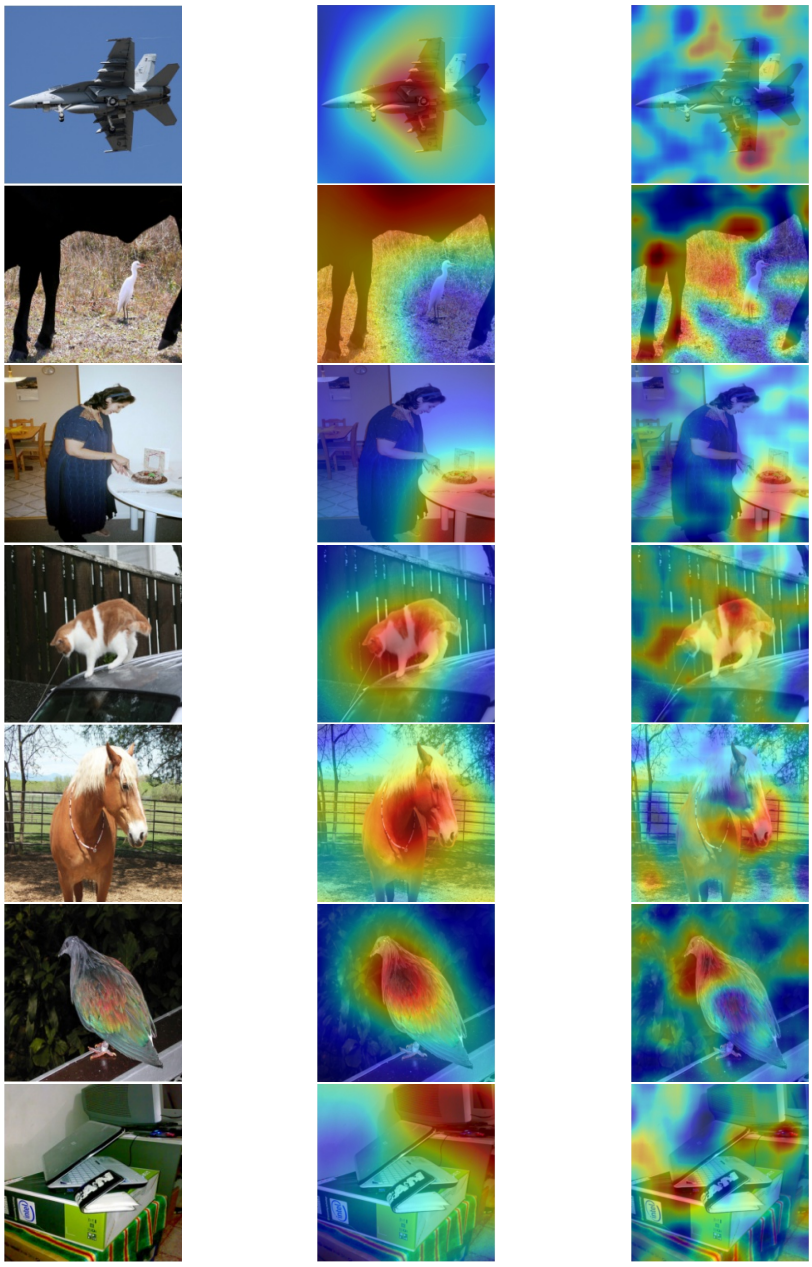
Table 5

Metric	Base.	Comp.	Stat.	p-val.
Deletion	no_flip	normal	271573.0	0.004182
	square	normal	297592.0	0.348812
	no_normalize	normal	70370.0	0.383979
F-meas.	no_flip	normal	403894.0	1.171645e-22
	square	normal	290082.0	8.650028e-01
	no_normalize	normal	75747.0	0.112701
Insertion	no_flip	normal	287112.0	9.170745e-01
	square	normal	242094.0	1.0
	no_normalize	normal	102515.0	1.439060e-18



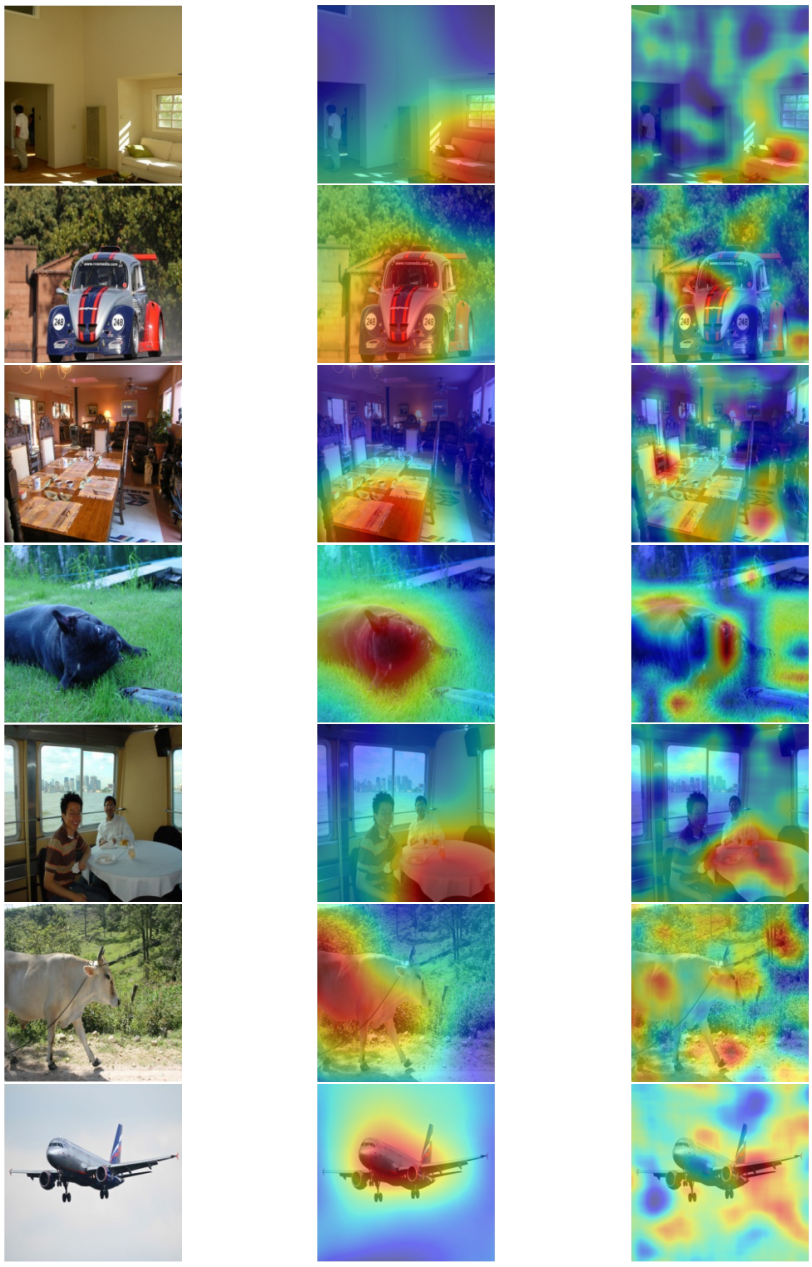
(a) Input (b) BOREx (c) RISE

Fig. 1: Examples of saliency maps that are successfully refined by BOREx. The labels used in each explanation are: “bird”, “motor bike”, “dining table”, “bird”, “bird”, “sheep”, and “bicycle”, from the first row.



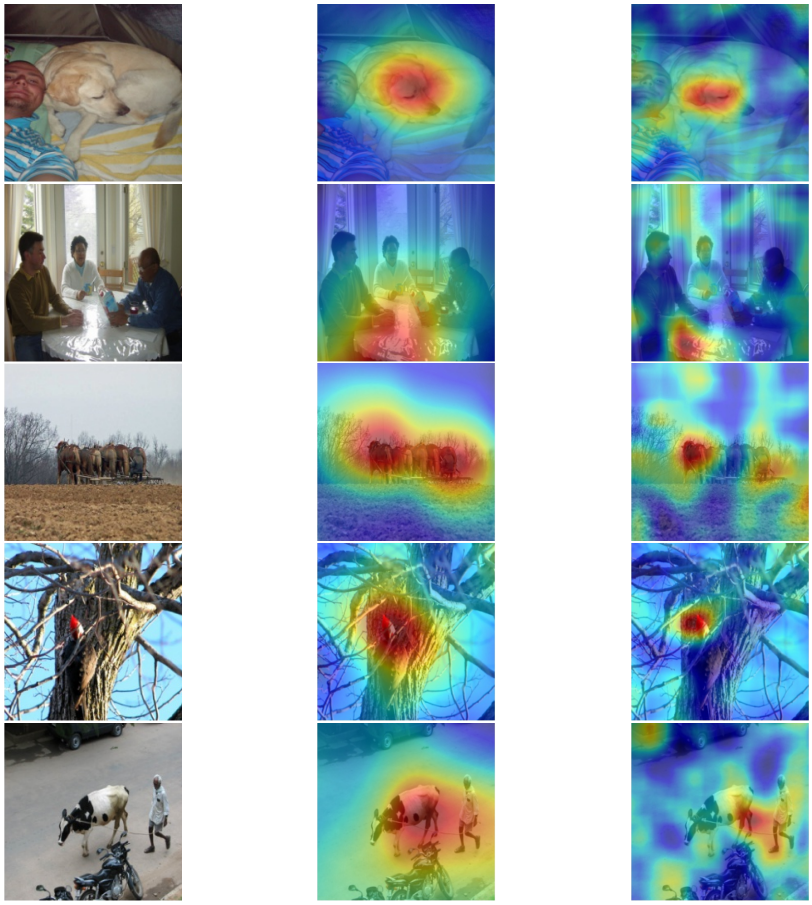
(a) Input (b) BOREx (c) RISE

Fig. 2: Examples of saliency maps that are successfully refined by BOREx. The labels used in each explanation are: “aeroplane”, “cow”, “dining table”, “cat”, “horse”, “bird”, and “TV monitor”, from the first row.



(a) Input (b) BORExp (c) RISE

Fig. 3: Examples of saliency maps that are successfully refined by BORExp. The labels used in each explanation are: “sofa”, “car”, “dining table”, “dog”, “dining table”, “cow”, and “aero plane”, from the first row.

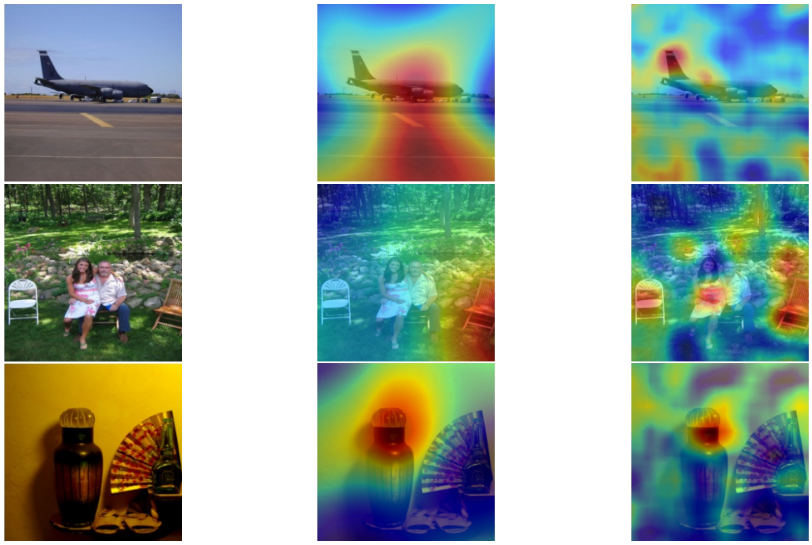


(a) Input (b) BOREx (c) RISE

Fig. 4: Examples of saliency maps that are successfully refined by BOREx. The labels used in each explanation are: “dog”, “dining table”, “horse”, “bird”, and “cow”, from the first row.

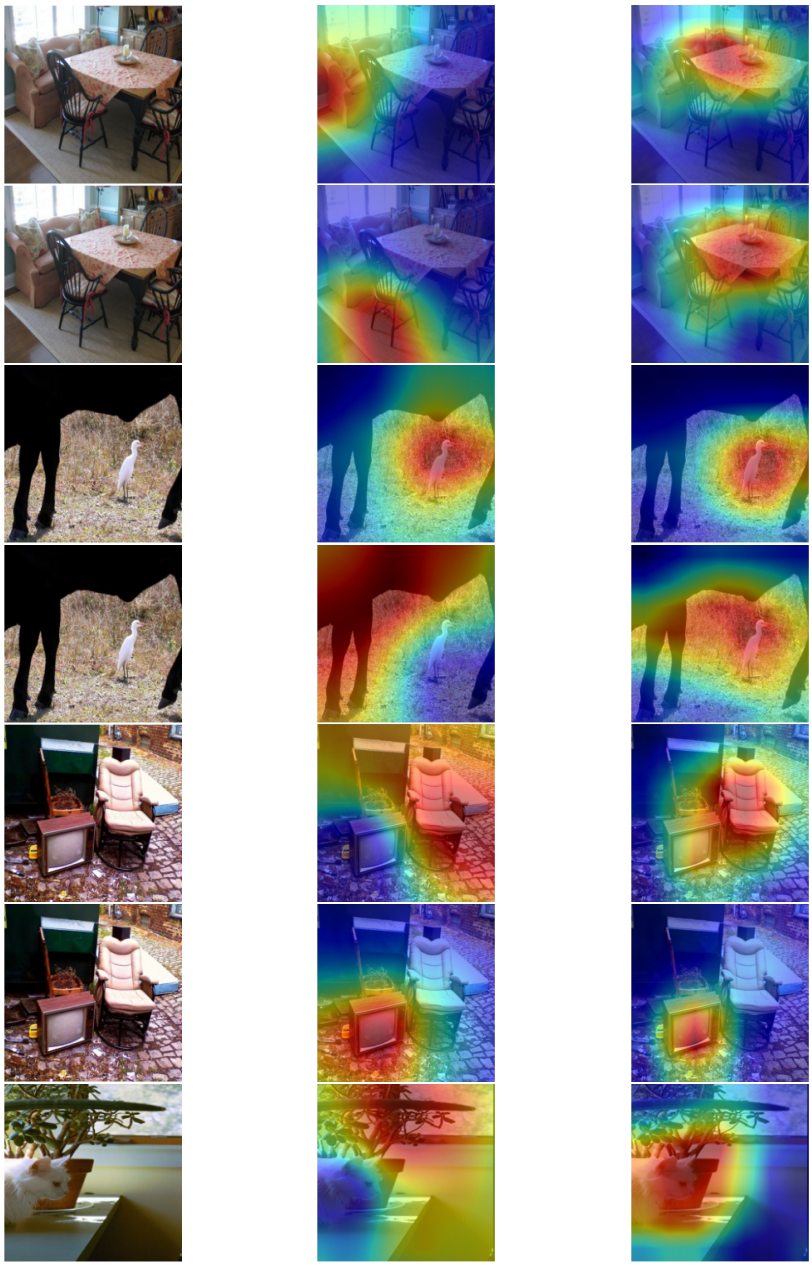
Table 6

Compared w/	Metric	Stat.	p-val.
RISE	F-meas.	63841.0	8.307474e-21
	ins.	65484.0	1.016050e-23
	del.	44608.0	0.887350
Grad-CAM++	F-meas.	16054.0	1.0
	ins.	2731.0	0.363604
	del.	54406.0	5.089895e-08



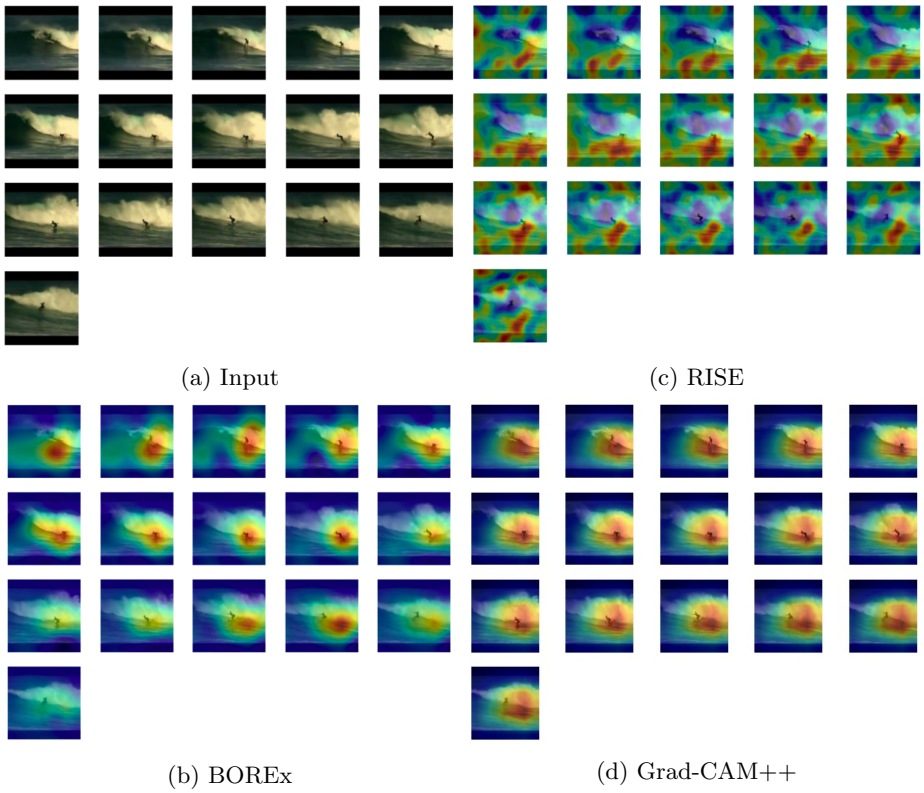
(a) Input (b) BOREx (c) RISE

Fig. 5: Examples of saliency maps that BOREx degraded the quantitative metric of the input image synthesized by RISE. The labels used in each explanation are: "warplane", "park bench", and "water bottle", from the first row.



(a) Input (b) BOREx (c) Grad-CAM++

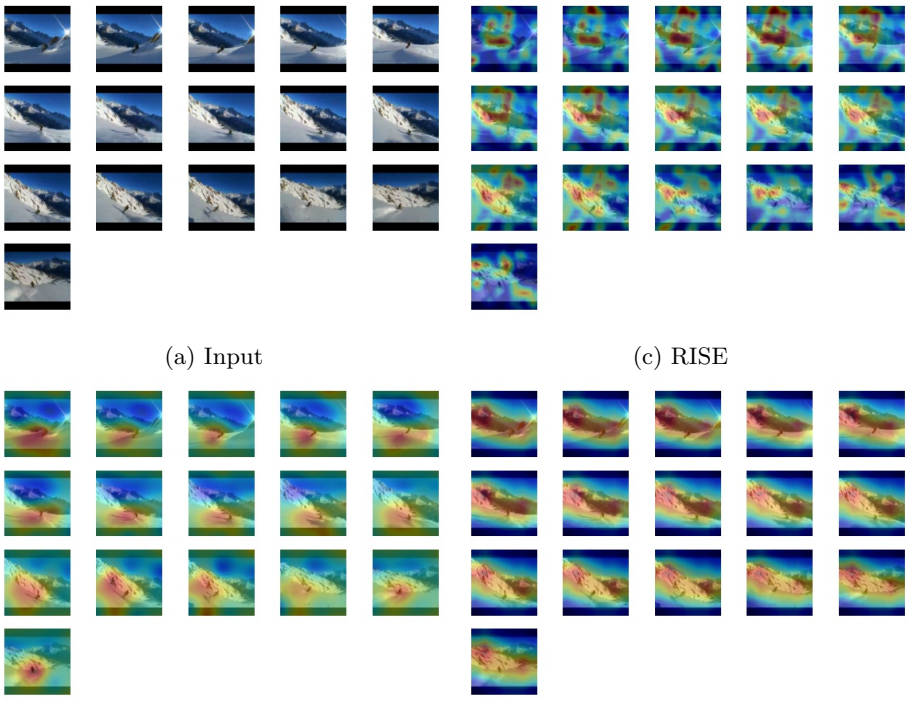
Fig. 6: Examples of saliency maps in which BOREx outperforms GradCAM++. The labels used in each explanation are: “sofa”, “chair”, “bird”, “cow”, “chair”, “TV monitor”, and “potted plant”, from the first row.



(a) Input (c) RISE

(b) BORExp (d) Grad-CAM++

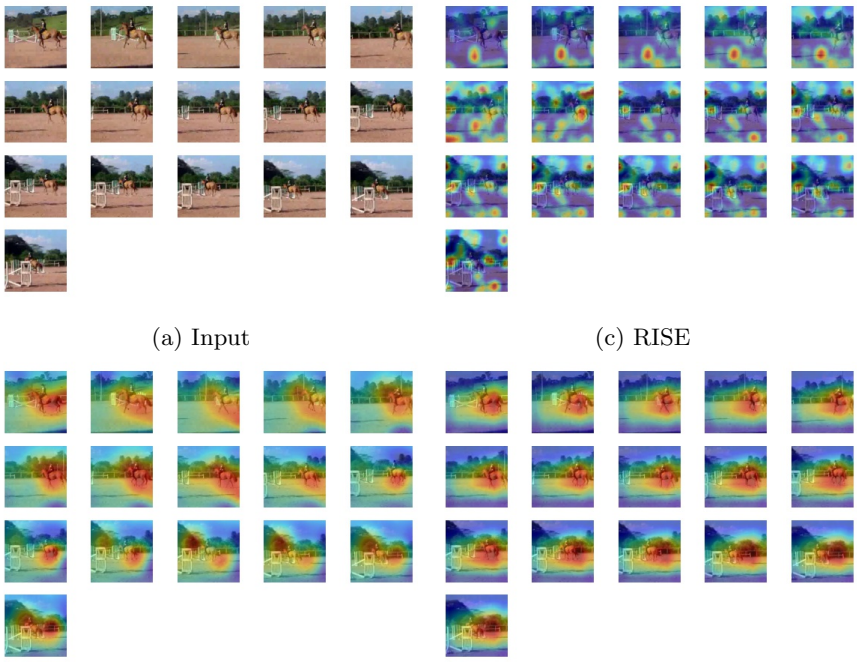
Fig. 7: Examples of saliency maps for video classifiers with label “surfing”.



(a) Input (c) RISE

(b) BOREx (d) Grad-CAM++

Fig. 8: Examples of saliency maps for video classifiers with label “skiing”.



(a) Input (c) RISE

(b) BORExp (d) Grad-CAM++

Fig. 9: Examples of saliency maps for video classifiers with label “horseriding”.