MFPP: Morphological Fragmental Perturbation Pyramid for Black-Box Model Explanations

Qing Yang, Xia Zhu, Jong-Kae Fwu, Yun Ye, Ganmei You and Yuan Zhu
{qing.y.yang, xia.zhu, jong-kae.fwu, yun.ye, ganmei.you, yuan.y.zhu}@intel.com

Introduction

Deep neural networks have made breakthroughs in various AI tasks and greatly changed many fields, such as computer vision and natural language processing. However, the lack of transparency of the DNN model has led to serious concerns about these DNN models given decision-making power in critical applications. The proposed method MFPP provides importance map on which area support each result in a CNN model prediction.

Definition

$$S_{I,\Phi}(\mu) \approx \frac{1}{E[M]\cdot N}\sum_{i=1}^{N}\Phi(I\odot M_i)\cdot M_i(\mu) \qquad (1)$$

where Ψ is image segmentation operation, and F_l is the output of image I from this operation:

$$F_l = \Psi(I) \qquad (2)$$

In this case, N is the total number of masks with different segmentation scales, $g(l)$ is the number of fragments in group l and L is the total number of groups.

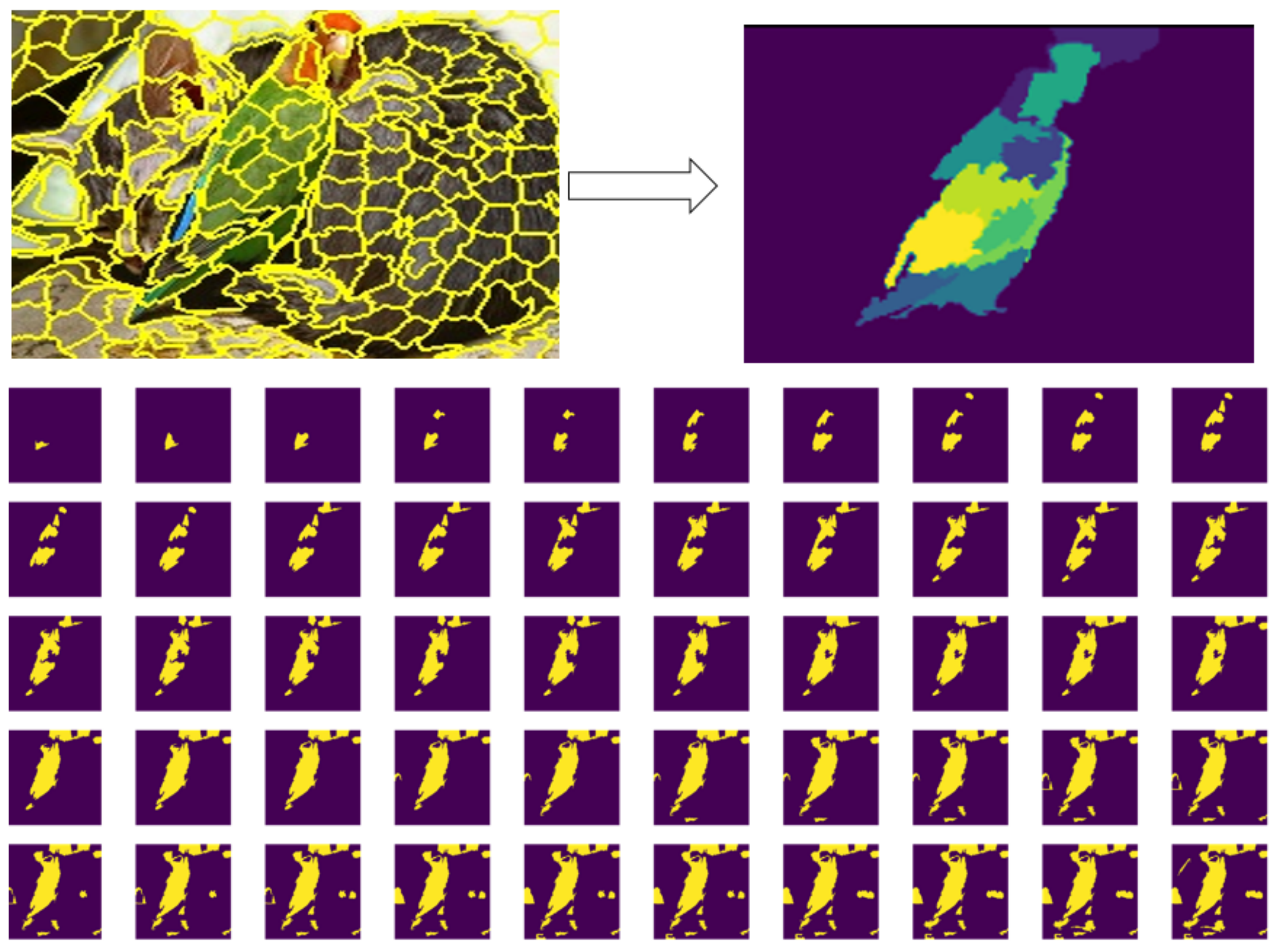
$$N = \sum_{l=1}^{L} g(F_l) \qquad (3)$$

Substituting N from (3) in (1)

$$S_{I,\Phi}(\mu) = \frac{1}{E[M]\cdot\sum_{l=1}^{L} g(F_l)}\sum_{l=1}^{L}\sum_{i=1}^{g(F_l)}\Phi(I\odot M_{l,i})\cdot M_{l,i}(\mu) \qquad (4)$$

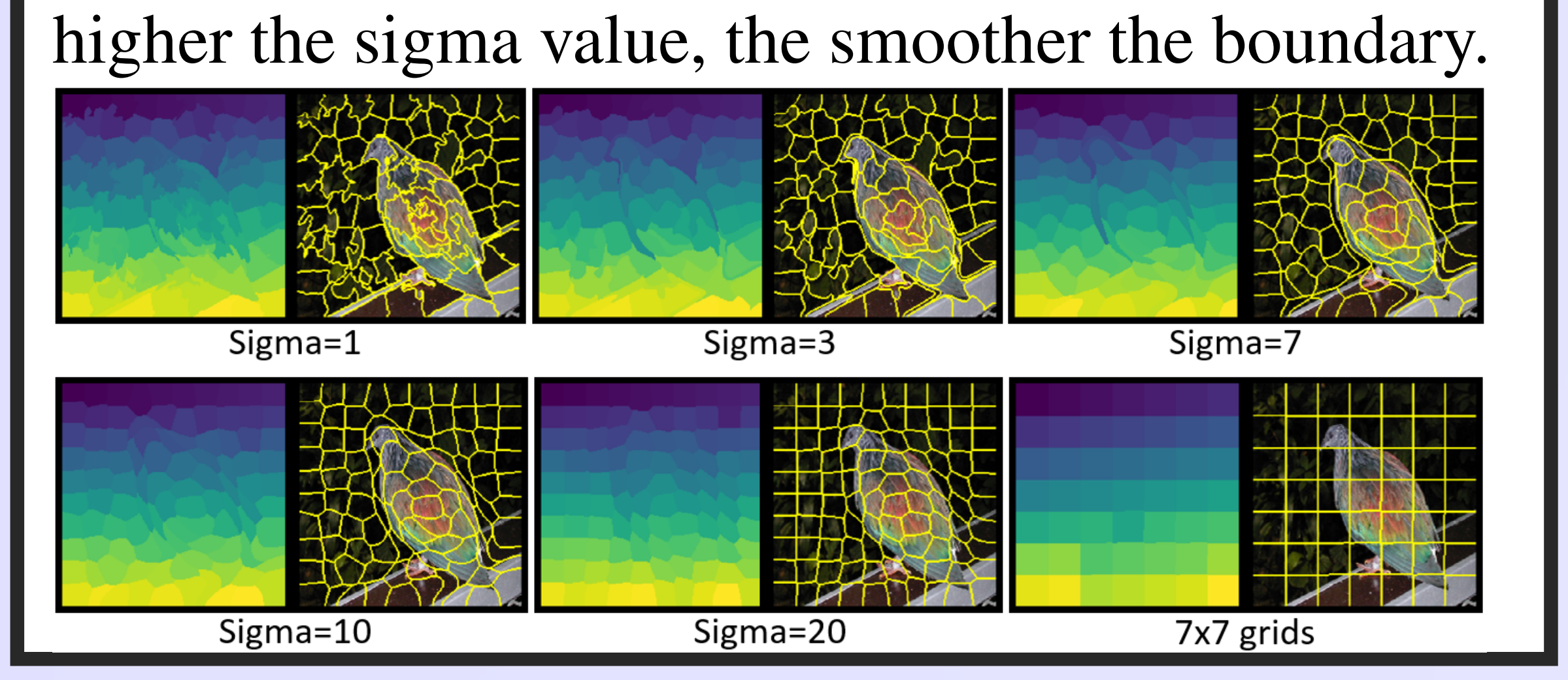
Study 1: Importance Distribution

Different part of an object also obtain different possibility score for its label.



Study 2: Morphological Degree

The sensitivity of MFPP to fragmentation morphological degree controls by sigma value of SLIC. The higher the sigma value, the smoother the boundary.



References

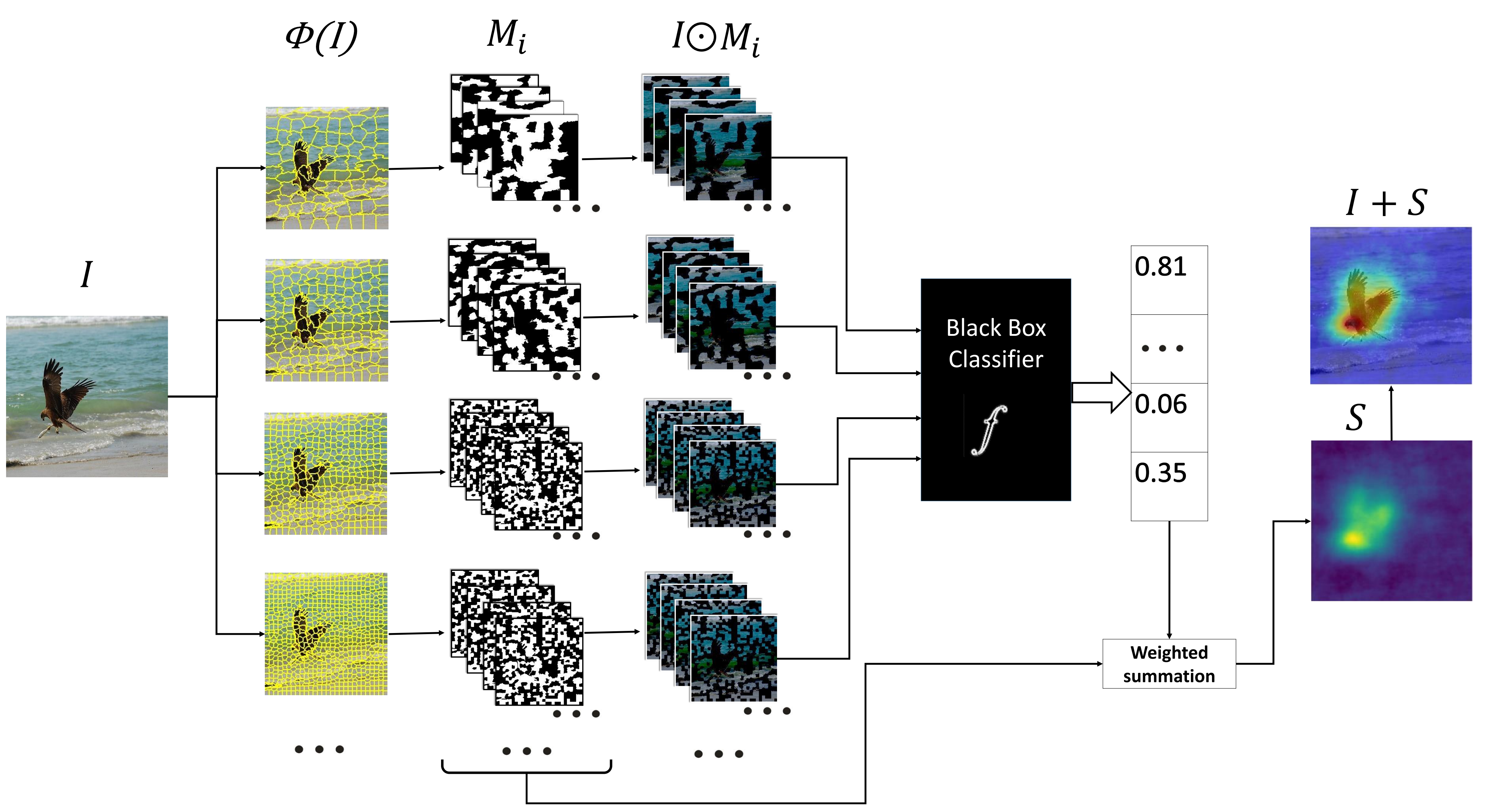
[1] Fong, Ruth and Patrick, Mandela and Vedaldi, Andrea: *Understanding deep networks via extremal perturbations and smooth masks*, Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2950–2958.

[2] Zhang, Jianming et al.*Top-down neural attention by excitation backprop*, International Journal of Computer Vision, vol. 126, no. 10, pp. 1084–1102, 2018.

[3] Petsiuk, Vitali and Das, Abir and Saenko, Kate : *Rise: Randomized input sampling for explanation of black-box models*. arXiv preprint arXiv:1806.07421.

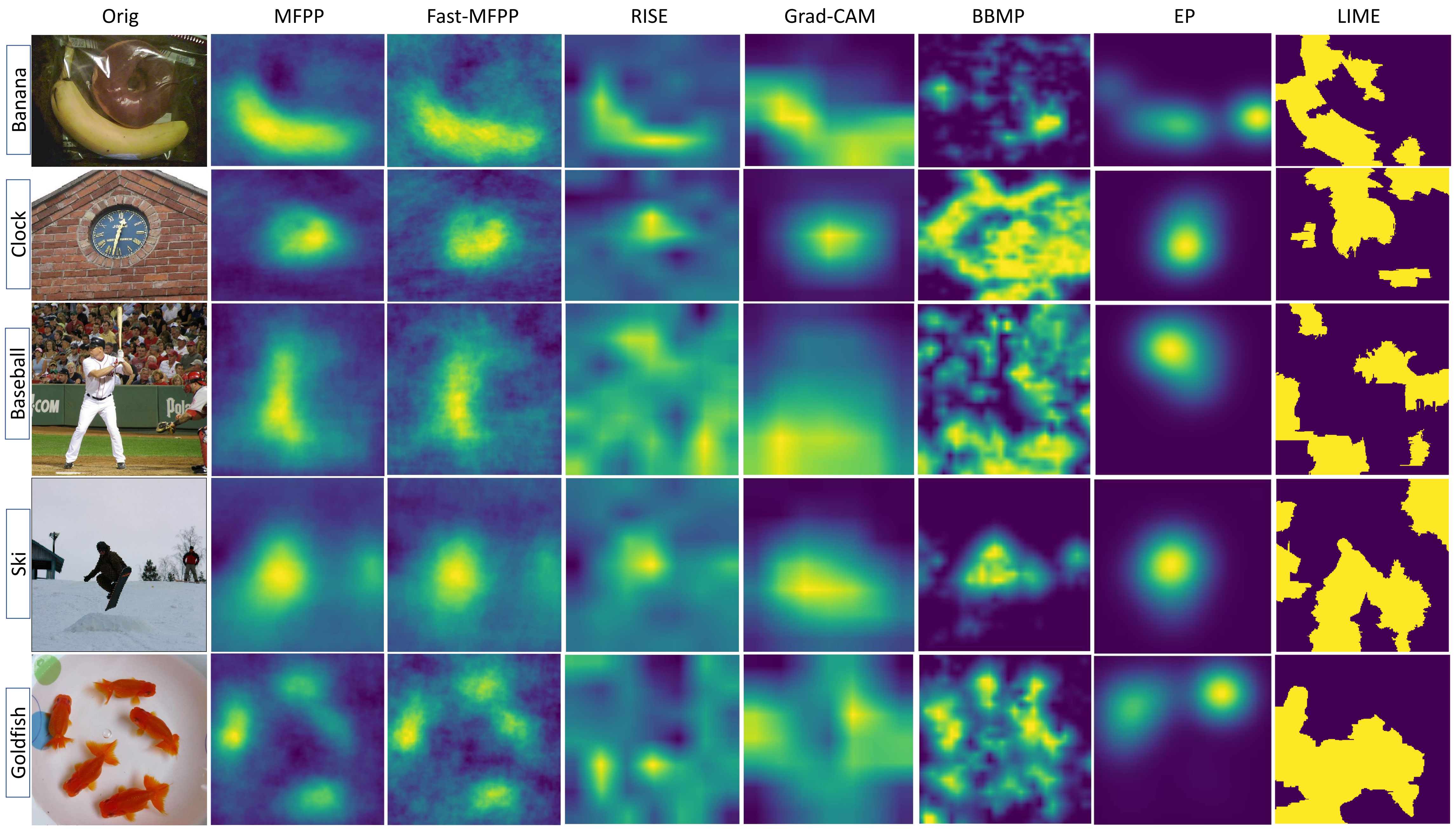
MFPP Methodology

This is the overview of MFPP methodology: The input image is sent to a segmentation algorithm to generate multiple-scale segmentation results. For each segmentation scale, masks are randomly generated and element-wise multiplied with input to obtain masked images. These masked images are fed to the black-box DNN model to obtain the saliency scores of the target classes. The scores are weighted summed across all masks. The final output is the saliency map.



Experiments : MFPP vs Previous works

Model explanation result of the proposed MFPP and its fast version and five other methods from left to right: RISE, Grad-CAM, BBMP, EP, LIME with VGG16 pretrained model on images from MS COCO2014 dataset. For intuitive visual evaluation, MFPP provides a more accurate and fine-grained saliency map than other competitive methods.



Pointing game experiments are adopted to evaluate the accuracy on the PASCAL VOC07 test set and COCO minival set . The result show that our proposed MFPP method benefits from morphological fragmentation and multiple perturbation layers. In terms of accuracy, it meets or exceeds the performance of the existing SOTA black box interpretation methods.

TABLE I: The Result of Pointing Game [30] on VOC2007 *test* and COCO2014 *minival* Dataset. Methods in Grey Color are for Black-box Model. EP's result on VOC07 is taken from [9].

Method	VOC07 Test		COCO14 MiniVal	
	VGG16	ResNet50	VGG16	ResNet50
Cntr.	69.6	69.6	27.6	27.6
Grad	76.3	72.3	37.4	35.4
DConv	67.5	68.6	30.5	30.2
Guid.	75.9	77.2	38.4	41.4
MWP	77.1	84.4	39.2	48.8
cMWP	79.9	90.7	49.8	57.4
Grad-CAM	86.6	90.4	54.0	57.0
RISE	86.4 ± 0.6	86.6 ± 1.0	51.1 ± 0.1	54.4 ± 0.4
EP	**88.0**	88.9	51.5 ± 0.1	56.1 ± 0.2
Fast-MFPP	86.1 ± 0.2	88.7 ± 0.4	50.6 ± 0.2	54.5 ± 0.3
MFPP	87.0 ± 0.1	**89.1 ± 0.6**	**52.0 ± 0.2**	**56.4 ± 0.2**

Pointing Game

$$Acc = \frac{Hits}{Hits+Misses}$$

TABLE II: The Benchmark for Average Processing Time for Single Sample Explanation on VOC07 Test Dataset.

Method	VOC07 Test		COCO14 MiniVal	
	VGG16	ResNet50	VGG16	ResNet50
LIME	39.8 ± 0.9	32.3 ± 1.1	35.6 ± 1.0	30.2 ± 1.4
RISE	17.0 ± 0.1	13.5 ± 0.1	**9.1 ± 0.3**	19.8 ± 0.4
EP	123.9 ± 0.2	72.1 ± 0.1	92.6 ± 3.5	75.9 ± 0.5
Fast-MFPP	**16.9 ± 0.1**	**6.7 ± 0.0**	**9.1 ± 0.3**	**10.0 ± 0.3**
MFPP	83.9 ± 0.1	32.6 ± 0.1	45.1 ± 0.3	50.1 ± 0.2