Improving Interpretability by Information Bottleneck Saliency Guided Localization

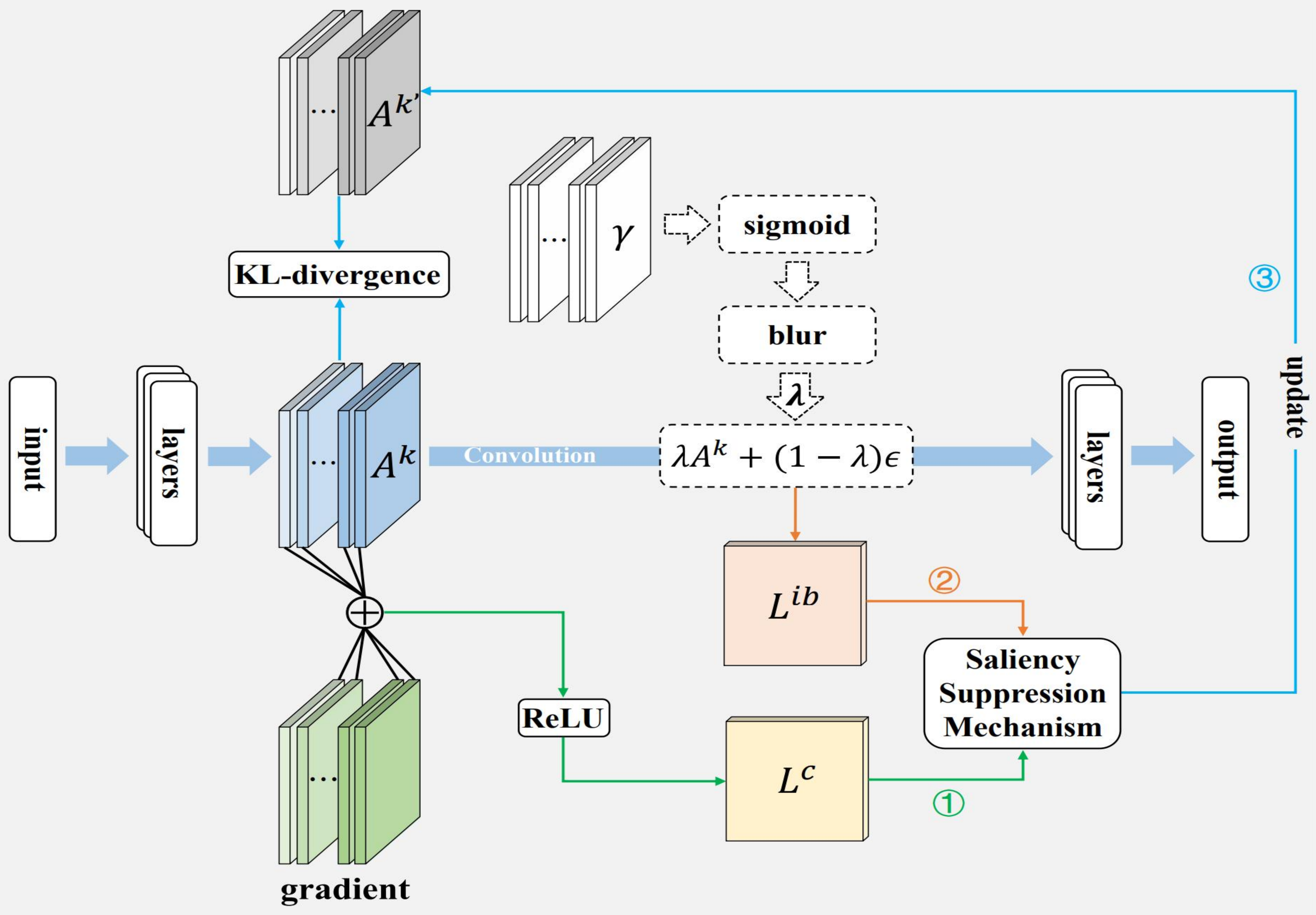
Hao Zhou, Keyang Cheng, Yu Si, Liuyang Yan

BMVC 2022

Motivation

- Deep Neural Networks are usually uninterpretable and untrustworthy, especially in high-risk domains;
- Current interpretable methods produce saliency maps that are often noisy or do not match human knowledge;
- The neural network only extracts features from the background of the image instead of the foreground, due to the noise and uncertainty in the dataset.

Method

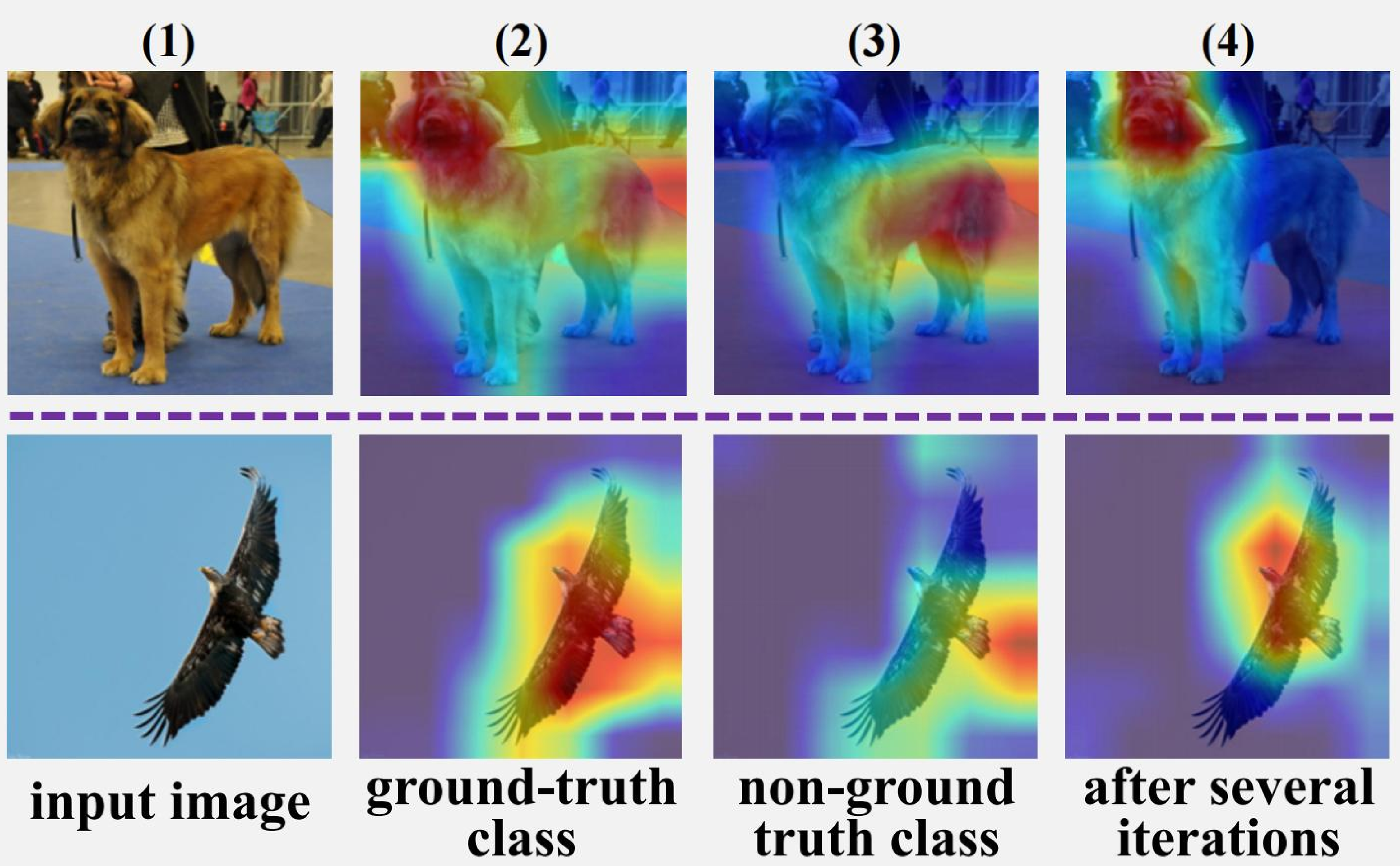


The proposed overview model for information bottleneck saliency-guided localization. step ① implements a gradient-based saliency map. step ② implements a saliency map based on information bottleneck attribution. step ③ implements an information bottleneck saliency map is used to guide model training.

Saliency Suppression Mechanism

we propose a new learning objective that incorporates the discrepancy between saliency maps as part of the learning process. The saliency map L^c of the ground-truth class c and the saliency map L^p of the class p with the highest probability are given, where the saliency map L^p comes from the non-ground truth class with the highest classification probability.

$$\mathcal{L}_{SS}\left(L^c, L^p\right) = \frac{\sum_{ij}\left[\min\left(L^c, L^p\right)\cdot Mask_r\right]}{\sum_{ij}\left(L^c + L^p\right)}$$



Information Bottleneck Guided Localization

As shown in Algorithm 1, the loss between the information bottleneck and the gradient-based saliency map is computed using our proposed saliency suppression mechanism to update the entire convolutional neural network.

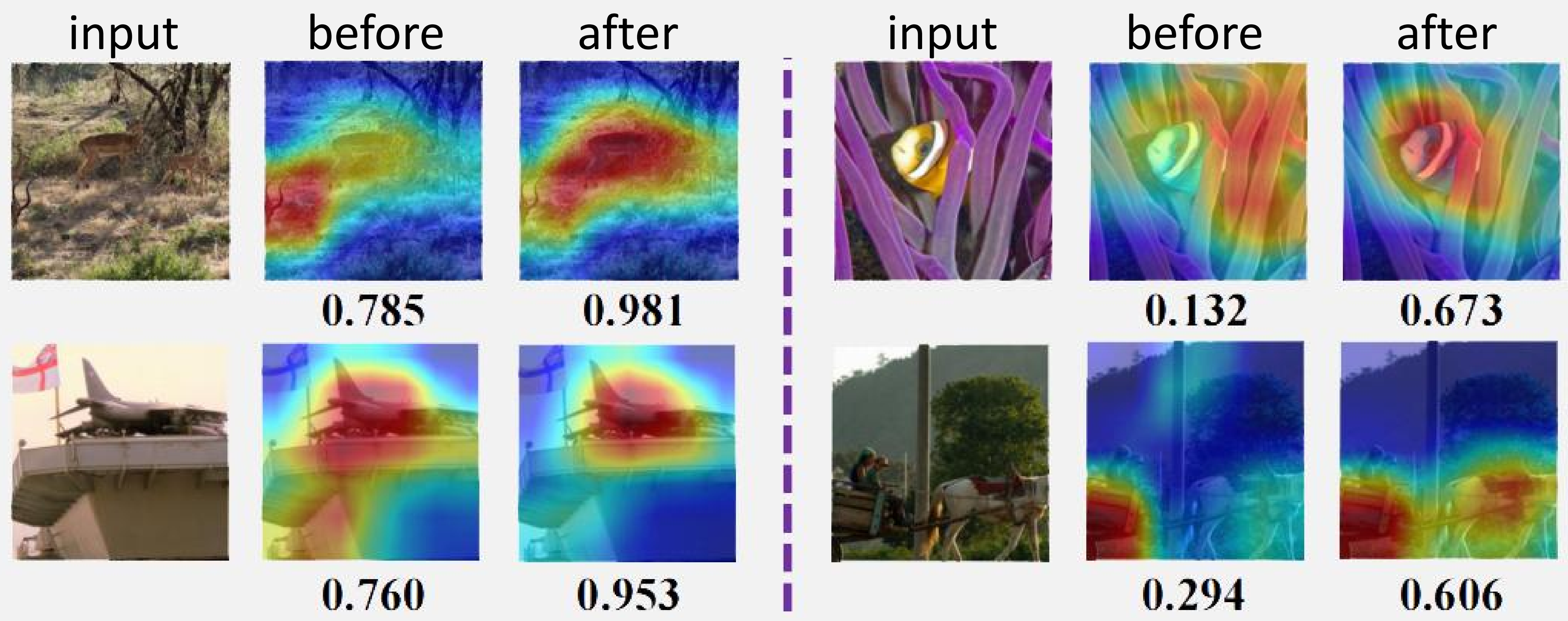
Algorithm 1 Saliency Guided Localization algorithm

Input: Image X_i, Class c, Sample size n of data set
Output: Model after saliency-guided localization

Initialization: Let A^k be the feature map of the last convolutional layer. y^c and y^p are the prediction scores for the target classes c and p, respectively.
for i in $[0, \dots, n-1]$ **do**
  Get the saliency map of class c and class p,
  $L^c, L^p \leftarrow \mathrm{ReLU}\left(\sum_k \frac{\partial y^c}{\partial A^k} A^k\right), \mathrm{ReLU}\left(\sum_k \frac{\partial y^p}{\partial A^k} A^k\right)$
  Saliency map L^{ib} is given based on information bottleneck attribution.
  if $c = p$ **then**
    $\underset{\omega}{\mathrm{minimize}} \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{L}_{CE}\left(L^c, L^{ib}\right) + \beta\mathcal{L}_{SS}\left(L^c, L^p\right)\right]$
  else
    $\underset{\omega}{\mathrm{minimize}} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{CE}\left(L^c, L^{ib}\right)$
  end if
  Update $A^k \rightarrow A^{k'}$ according to the above loss function.
  $\underset{\theta}{\mathrm{minimize}} \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{L}\left(f_\theta\left(X_i\right), y_i\right) + \mu D_{KL}\left(A^k \| A^{k'}\right)\right]$
  Update model parameters.
end for

Experiments

Guided Localization

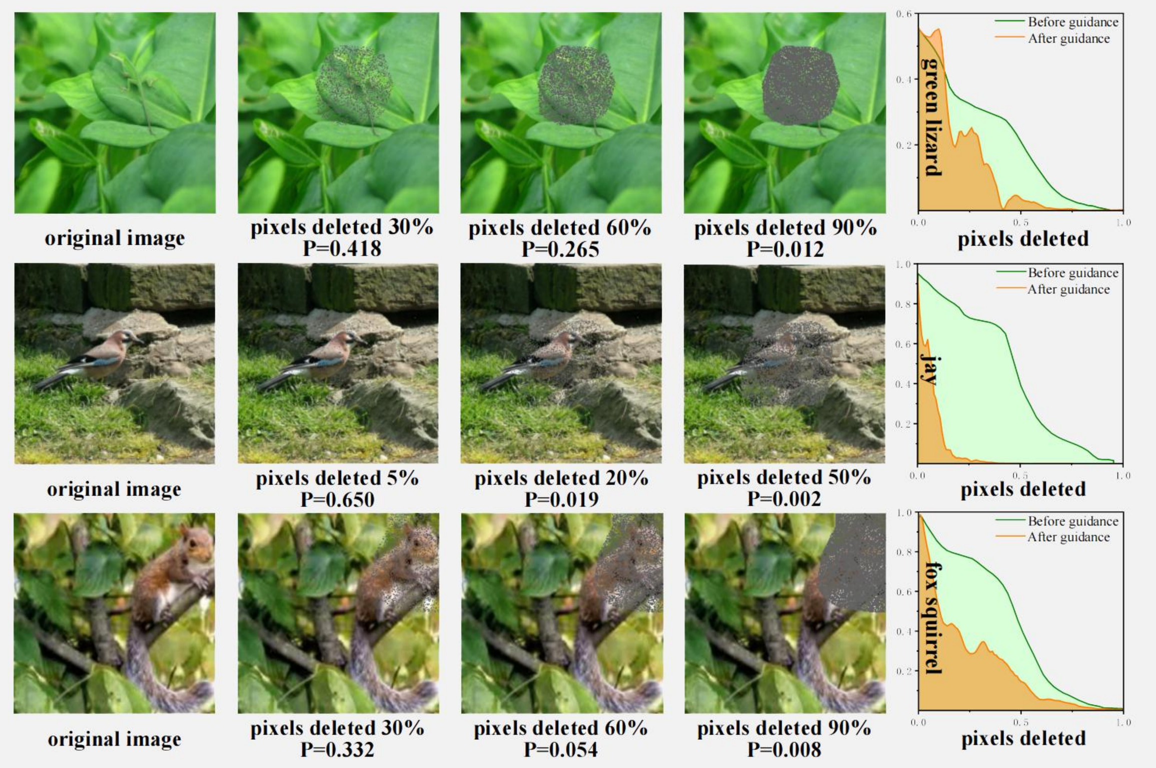


This approach allows the model's saliency map to converge on the saliency map of the information bottleneck and away from the saliency map of the non-ground truth.

Quantitative Evaluations

Metric	EBPG	mIoU	Bbox
Grad CAM [13]	60.08	32.16	**60.25**
Grad CAM++ [2]	47.78	30.16	58.66
Extremal Perturbation [4]	63.24	26.29	52.34
RISE [10]	32.86	27.40	55.55
Score CAM [19]	35.56	31.0	60.02
Integrated Gradient [16]	40.62	15.41	34.79
FullGrad [15]	39.55	20.20	44.94
Ours method	**65.07**	**32.74**	58.88

Metric	AD(%)	AI(%)
Grad CAM [13]	35.80	36.58
Grad CAM++ [2]	41.77	32.15
Extremal Perturbation [4]	39.38	34.27
RISE [10]	39.77	37.08
Score CAM [19]	35.36	37.08
Integrated Gradient [16]	66.12	24.24
FullGrad [15]	65.99	25.36
Ours method	**33.79**	**39.26**



Contributions

1. A novel interpretable method based on **information bottleneck saliency-guided localization** is proposed, which modifies the saliency map of the model to improve interpretability from the perspective of information theory.
2. We propose a **saliency suppression mechanism** that constrains the focus between ground-truth and non-ground truth saliency maps to reduce saliency from non-ground truth classes.

This work was supported by National Natural Science Foundation of China [61972183] and Jiangsu Province Science and Technology Program [BE2022781].