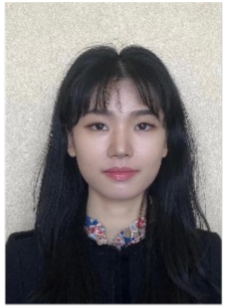
Less is More: Fewer Interpretable Region via Submodular Subset Selection

Ruoyu Chen

Hua Zhang

Siyuan Liang



Jingzhi Li

Xiaochun Cao

R. Chen's Homepage

WeChat

Paper

Code

Interpretable AI

Interpretation can improve human understandability, discover the errors made by the model, even can help revise and improve the model performance, and so on.

Interpretating high-performance deep learning models.

Universal interpretable method, with no need to pay attention to the model architecture itself, can be easily scaled to large models.

This work:

- design a more accurate and universal post-hoc attribution method
- discover what causes the model to make incorrect decisions

Image Attribution

The main objective in attribution techniques is to highlight the discriminating variables for decision-making.

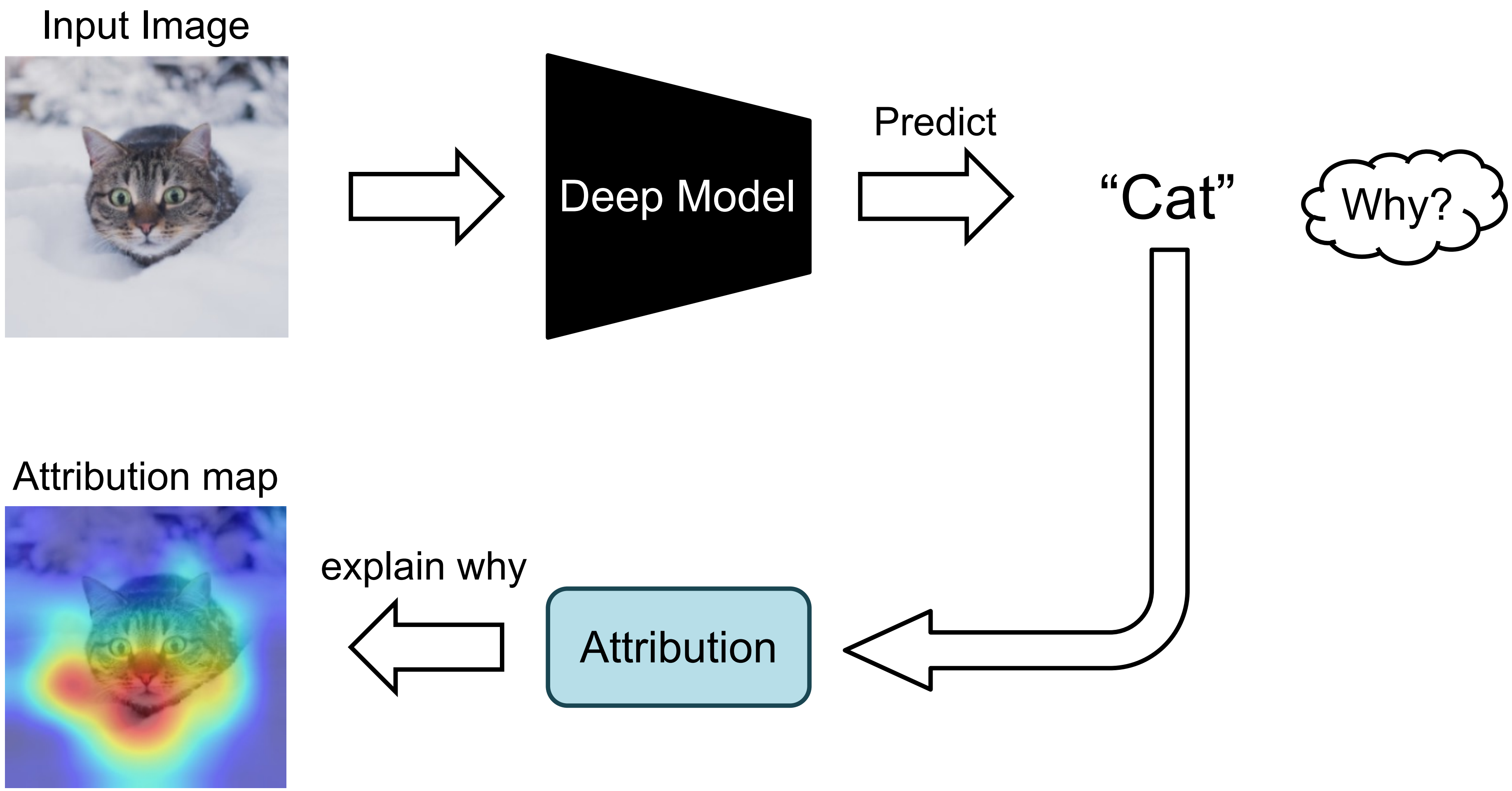
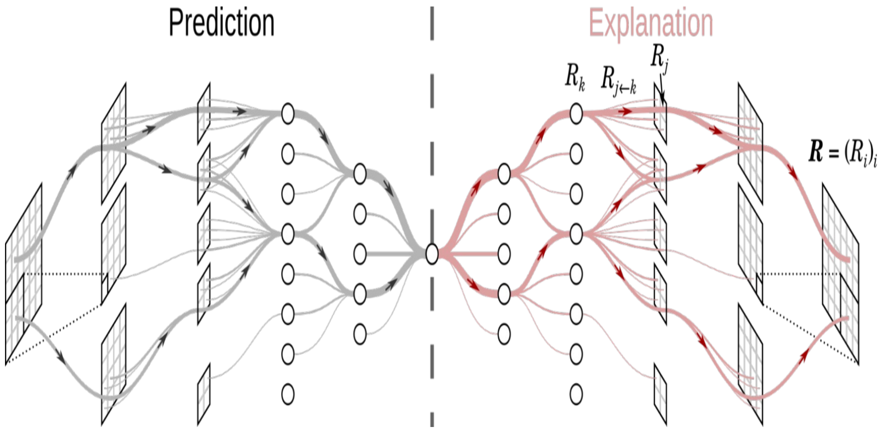


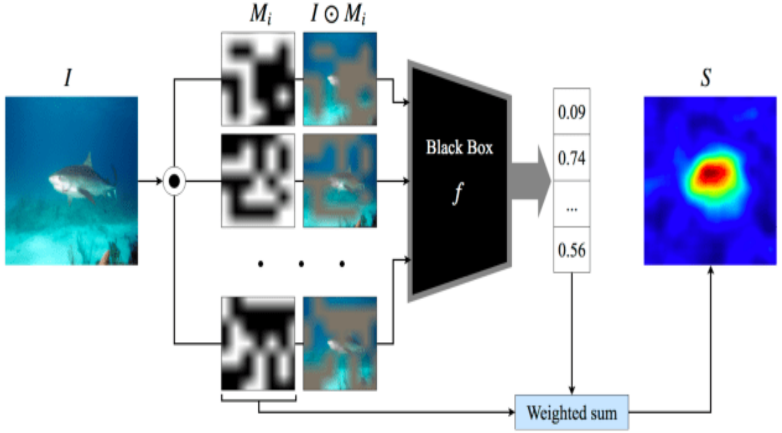
Image Attribution



Based on inner propagation, activation, or gradient

SHAP

Based on sharpley value estimation



Based on perturbation

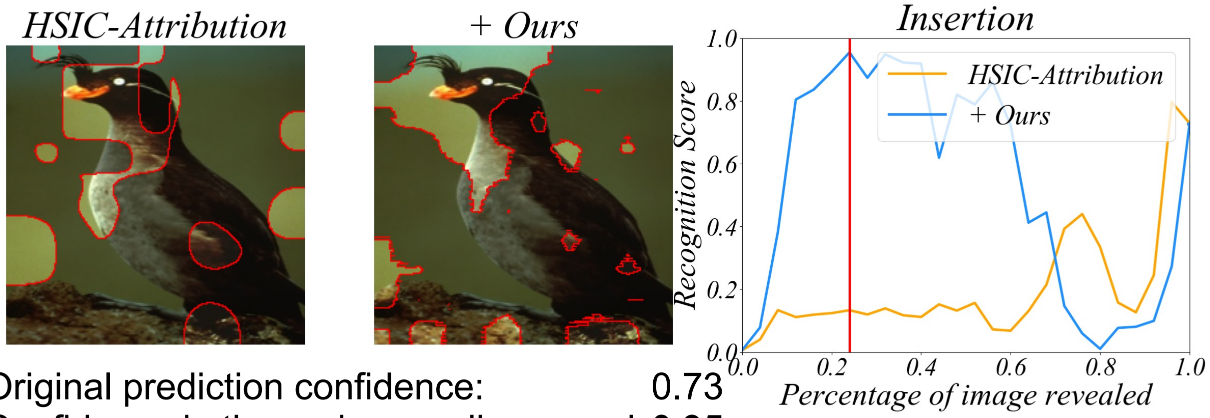
Challenge in Attribution

- Existing attribution methods generate *inaccurate small regions* thus misleading the direction of correct attribution.

Input Image $\mathbf{I}$

Attribution Map $\mathcal{A}$

Sample correctly predicted by the model

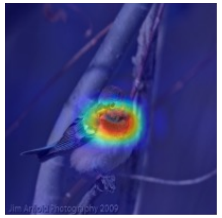


Original prediction confidence: 0.73
Confidence in the region we discovered: **0.95**

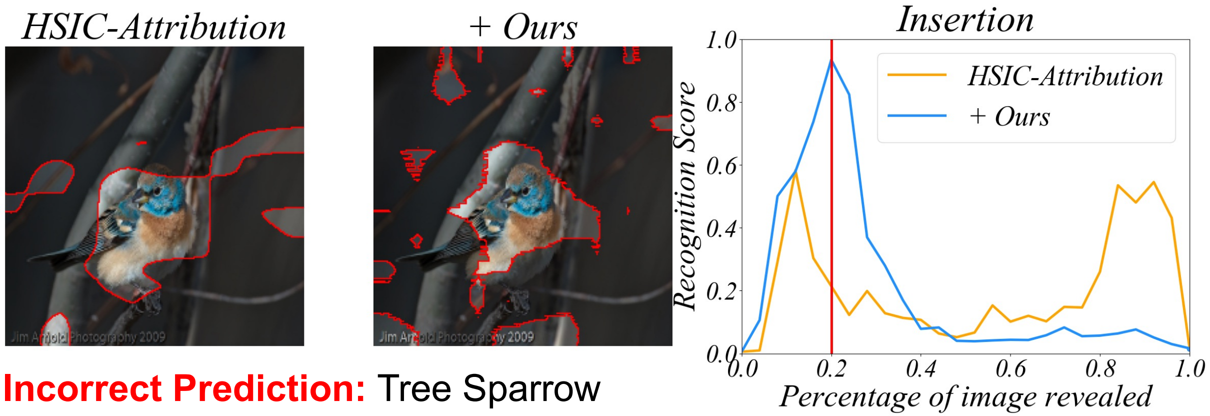
- They also can't produce good attribution results for samples with *wrong predictions*.

Input Image $\mathbf{I}$

Attribution Map $\mathcal{A}$



Sample incorrectly predicted by the model



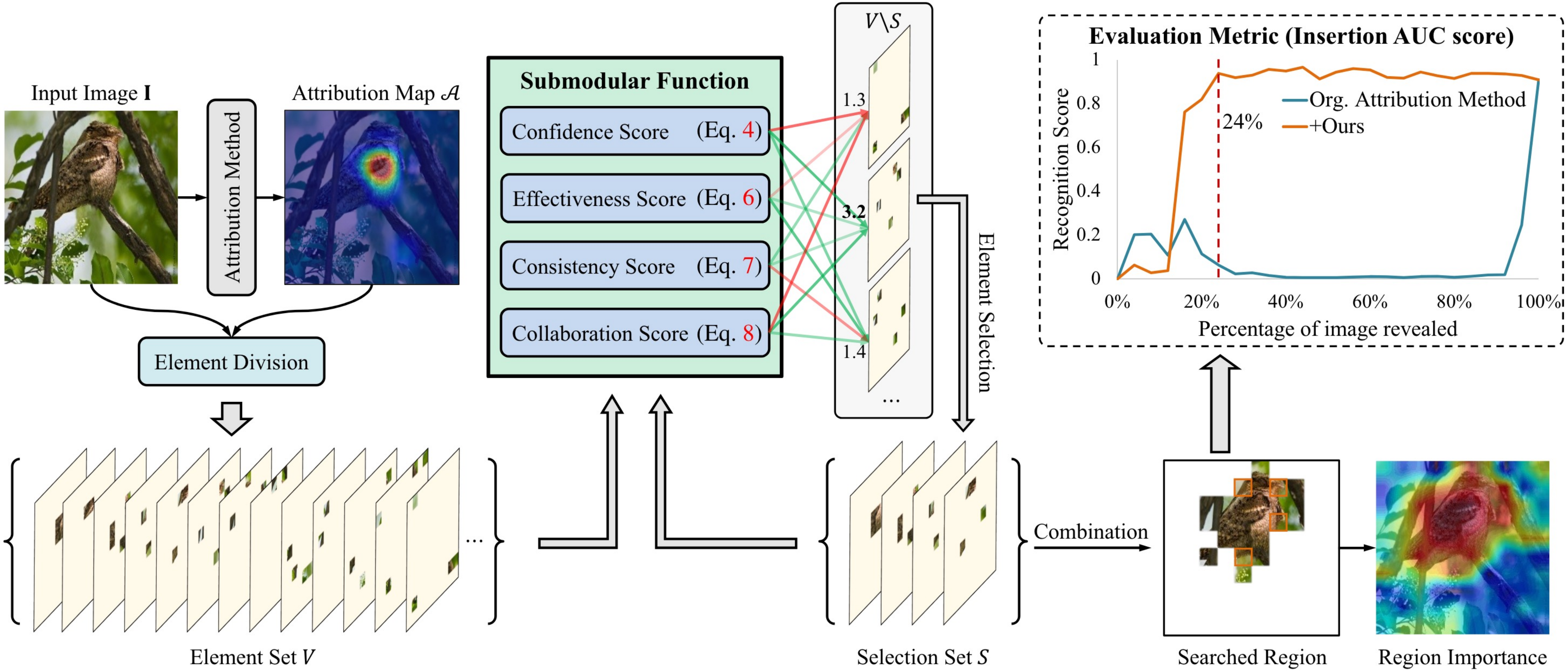
Incorrect Prediction: Tree Sparrow
Ground Truth: Lazuli Bunting

Our Solution

Divide the image into a set of small sub-regions and ranking the sub-regions according to their importance.

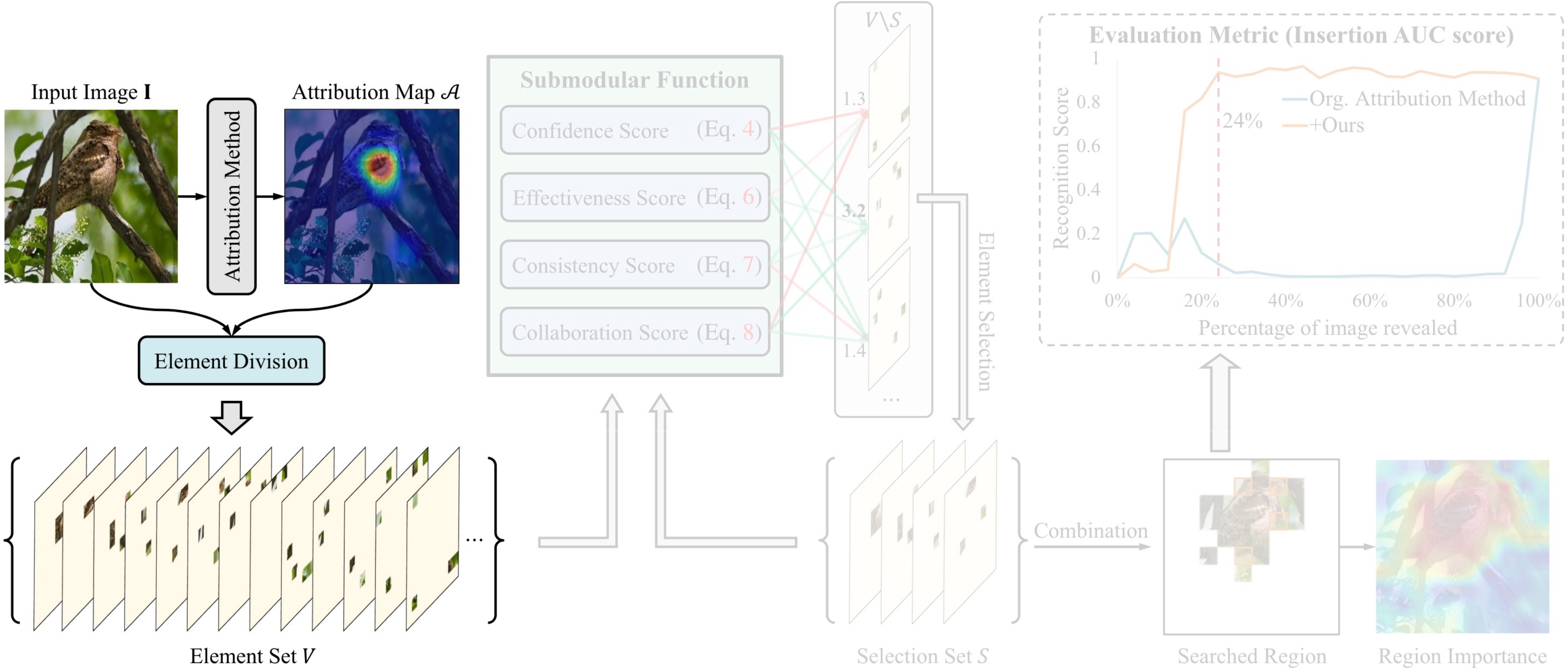
- Reformulate the attribution problem as a *submodular subset selection problem*;
- Employ regional *search* to expand the sub-region set to *alleviate the insufficient dense of the attribution region*;
- A novel *submodular mechanism* is constructed to *limit the search for regions with wrong class responses*.

Method



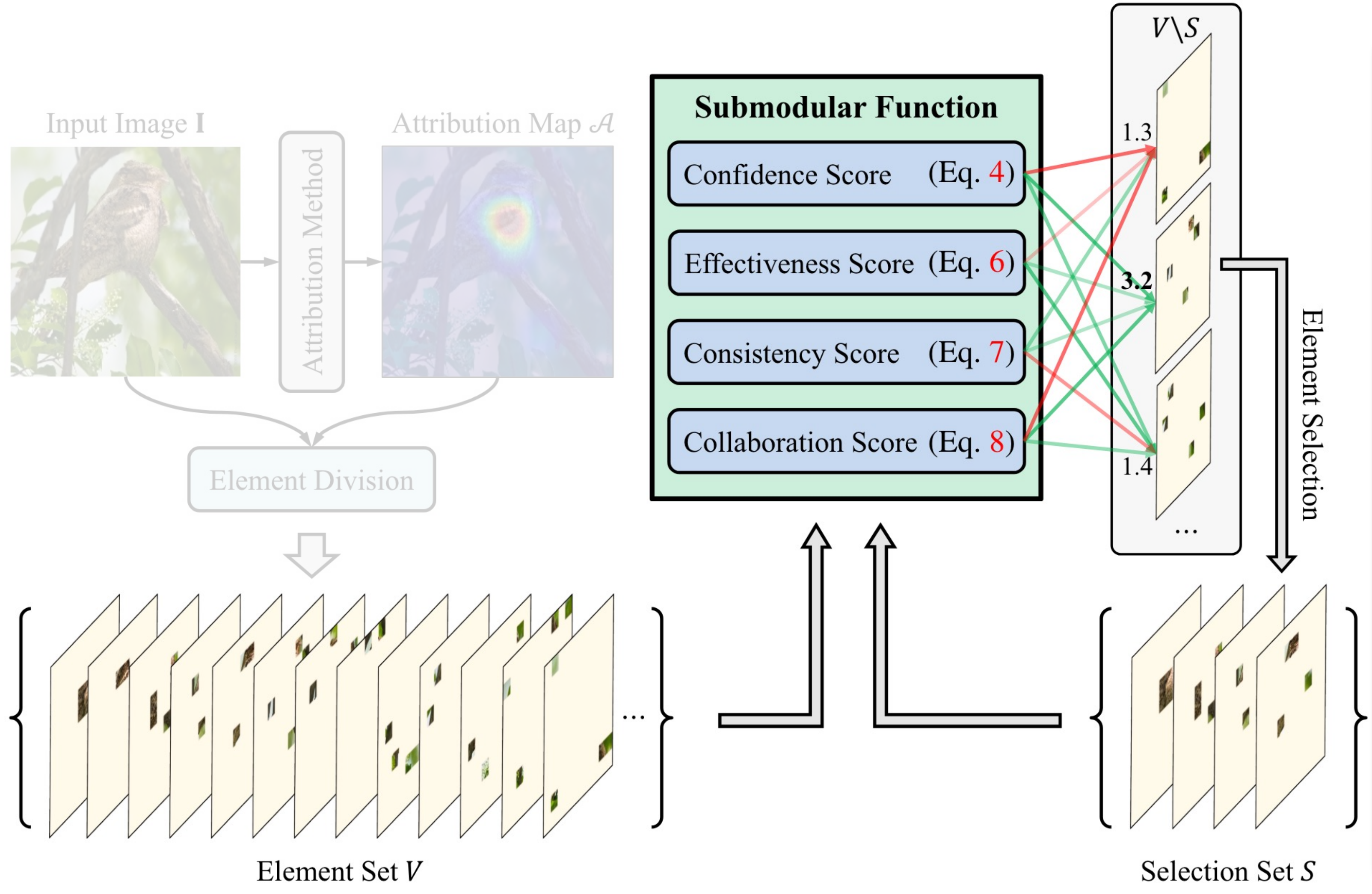
Input Image I
Attribution Method
Attribution Map $\mathcal{A}$
Element Division
Submodular Function
Confidence Score (Eq. 4)
Effectiveness Score (Eq. 6)
Consistency Score (Eq. 7)
Collaboration Score (Eq. 8)
$V \backslash S$
1.3
3.2
1.4
...
Element Selection
Evaluation Metric (Insertion AUC score)
Recognition Score
1
0.8
0.6
0.4
0.2
0
0%
20%
40%
60%
80%
100%
24%
Org. Attribution Method
+Ours
Percentage of image revealed
Element Set V
Selection Set S
Combination
Searched Region
Region Importance

Method



1. Sub-Region Division

Method



Designed Submodular Function

Confidence Score (*Improve credibility*):

$$s_{\text{conf.}}(\mathbf{x}) = 1 - \frac{K}{\sum_{k_c}^{K}(e_{k_c}+1)}, \qquad \text{(Eq. 4)}$$

Effectiveness Score (*Improve diversity*):

$$s_{\text{eff.}}(S) = \sum_{s_i \in S} \lim_{s_j \in S, s_i \neq s_j} \text{dist}\left(F(s_i), F(s_j)\right), \qquad \text{(Eq. 6)}$$

Consistency Score (*Improve semantic consis.*):

$$s_{\text{cons.}}(S, \boldsymbol{f}_s) = \frac{F\left(\sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) \cdot \boldsymbol{f}_s}{\left\| F\left(\sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) \right\| \|\boldsymbol{f}_s\|}, \qquad \text{(Eq. 7)}$$

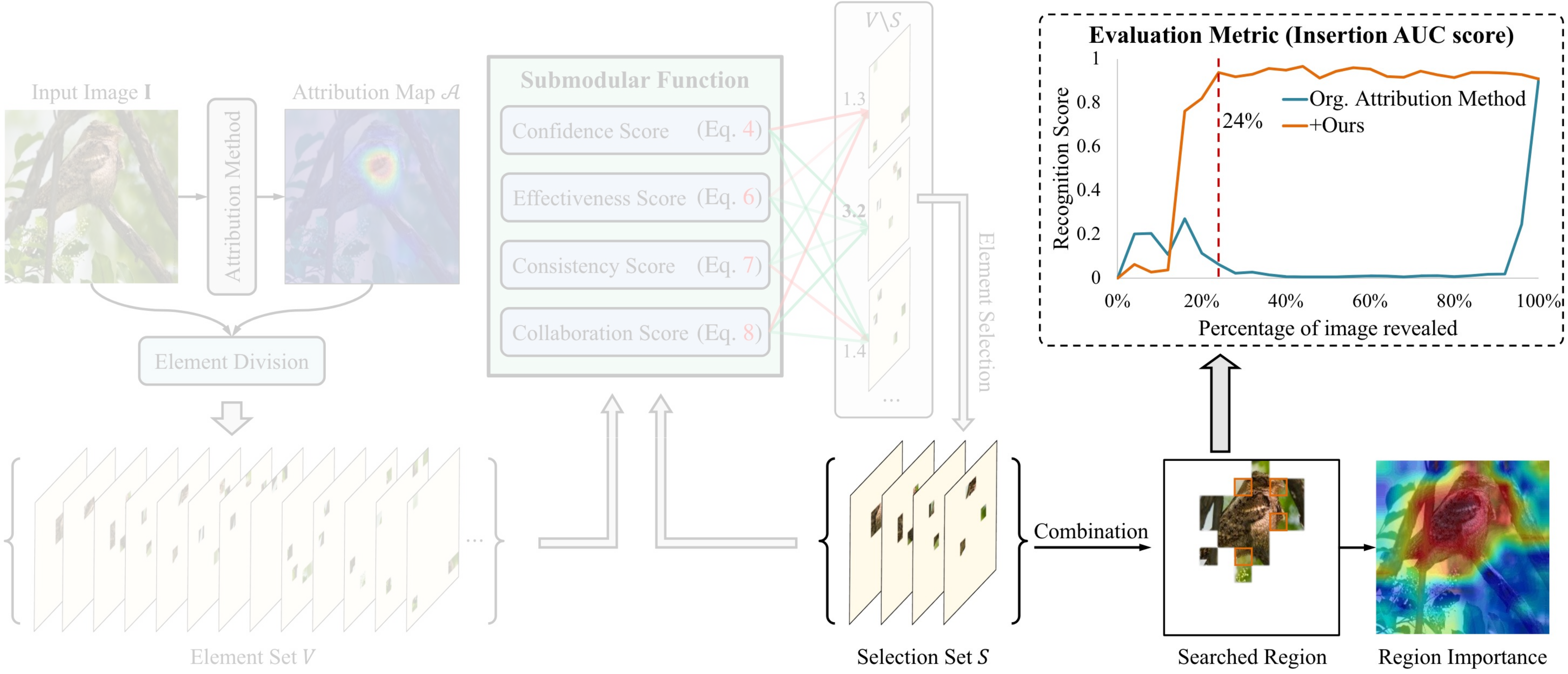
Collaboration Score (*Improve collective effect*):

$$s_{\text{colla.}}(S, \mathbf{I}, \boldsymbol{f}_s) = 1 - \frac{F\left(\mathbf{I} - \sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) \cdot \boldsymbol{f}_s}{\left\| F\left(\mathbf{I} - \sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) \right\| \|\boldsymbol{f}_s\|}, \qquad \text{(Eq. 8)}$$

1. Sub-Region Division

2. Sub-Region Selection

Method

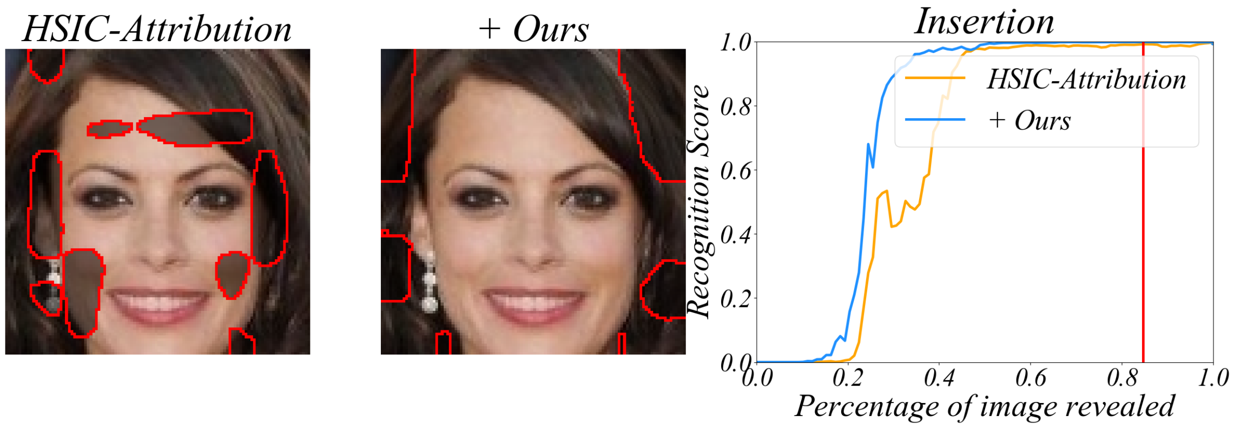


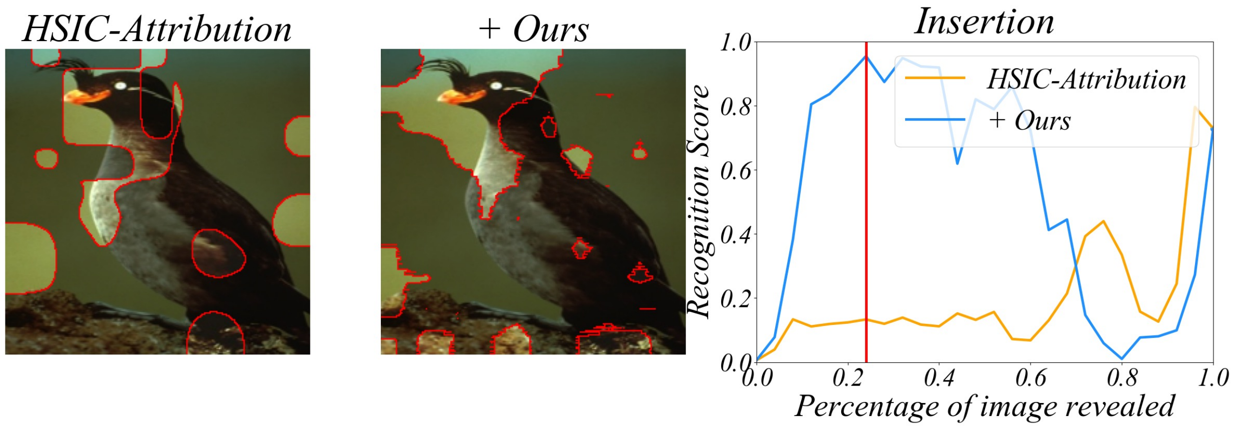
1. Sub-Region Division

2. Sub-Region Selection

3. Combination and Evaluation

Advanced Attribution Results





Use fewer image region but get higher prediction confidence.

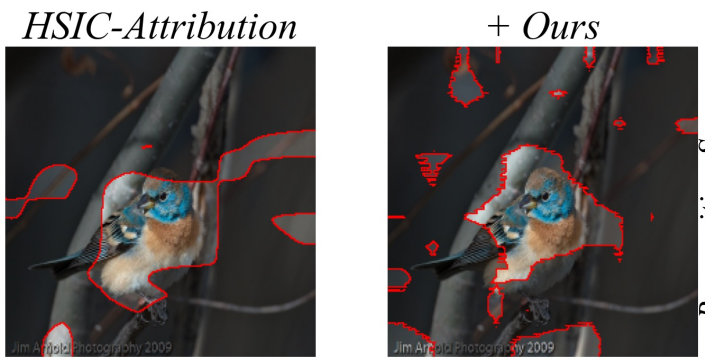
Table 1: Deletion and Insertion AUC scores on the Celeb-A, VGG-Face2, and CUB-200-2011 validation sets.

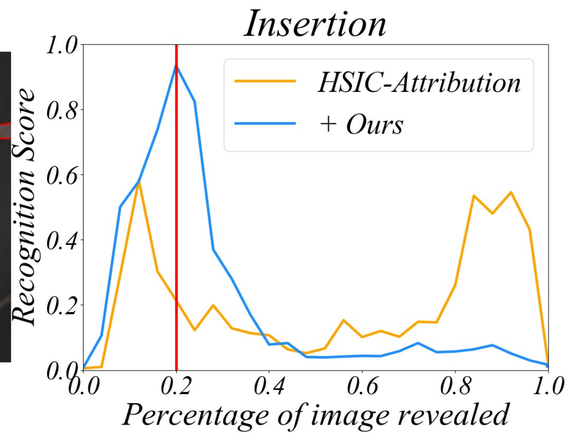
Method	Celeb-A		VGGFace2		CUB-200-2011	
	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)
Saliency (Simonyan et al., 2014)	0.1453	0.4632	0.1907	0.5612	0.0682	0.6585
Saliency (w/ ours)	**0.1254**	**0.5465**	**0.1589**	**0.6287**	**0.0675**	**0.6927**
Grad-CAM (Selvaraju et al., 2020)	0.2865	0.3721	0.3103	0.4733	0.0810	0.7224
Grad-CAM (w/ ours)	**0.1549**	**0.4927**	**0.1982**	**0.5867**	**0.0726**	**0.7231**
LIME (Ribeiro et al., 2016)	0.1484	0.5246	0.2034	0.6185	0.1070	0.6812
LIME (w/ ours)	**0.1366**	**0.5496**	**0.1653**	**0.6314**	**0.0941**	**0.6994**
Kernel Shap (Lundberg & Lee, 2017)	0.1409	0.5246	0.2119	0.6132	0.1016	0.6763
Kernel Shap (w/ ours)	**0.1352**	**0.5504**	**0.1669**	**0.6314**	**0.0951**	**0.6920**
RISE (Petsiuk et al., 2018)	0.1444	0.5703	0.1375	0.6530	0.0665	0.7193
RISE (w/ ours)	**0.1264**	**0.5719**	**0.1346**	**0.6548**	**0.0630**	**0.7245**
HSIC-Attribution (Novello et al., 2022)	0.1151	0.5692	0.1317	0.6694	0.0647	0.6843
HSIC-Attribution (w/ ours)	**0.1054**	**0.5752**	**0.1304**	**0.6705**	**0.0613**	**0.7262**

Deletion: ***4.9%*** improvement

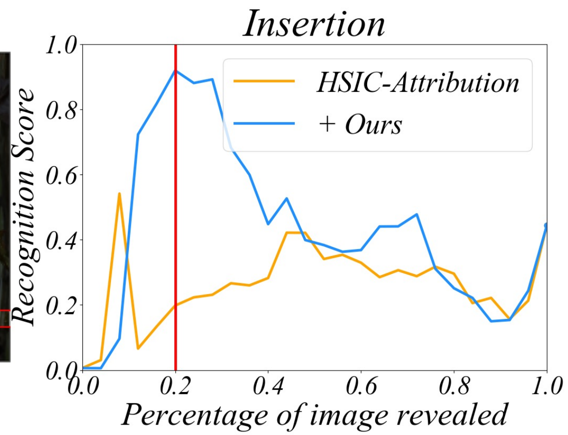
Insertion: ***2.5%*** improvement

Debugging Model Prediction Errors

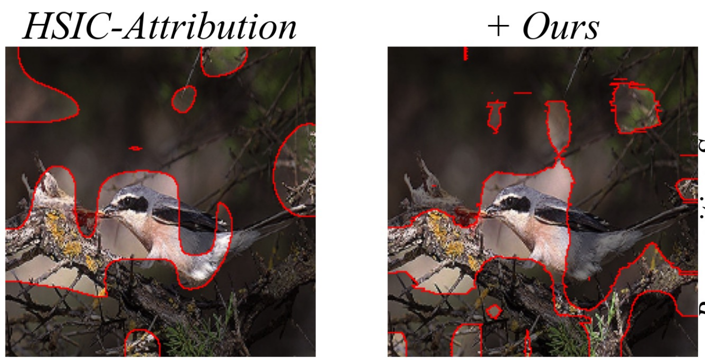


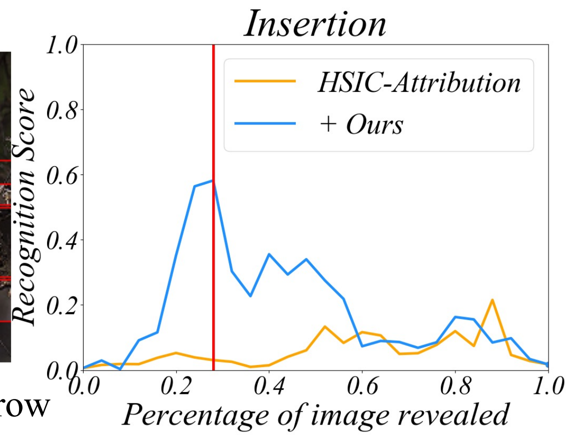


Incorrect Prediction: Tree Sparrow
Ground Truth: Lazuli Bunting

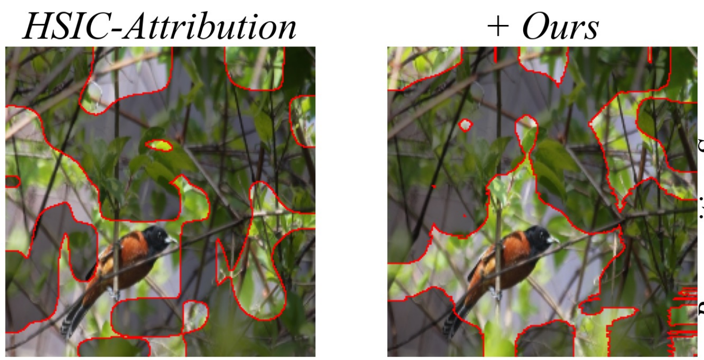


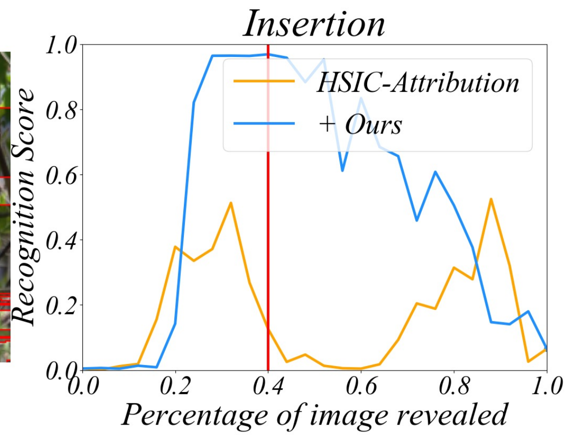
Incorrect Prediction: Tree Sparrow
Ground Truth: Chipping Sparrow





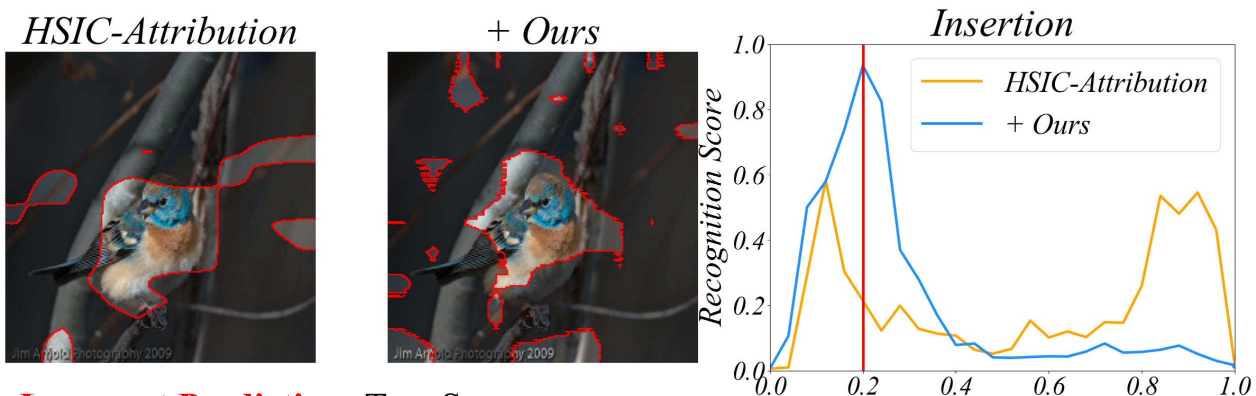
Incorrect Prediction: White Crowned Sparrow
Ground Truth: Great Grey Shrike



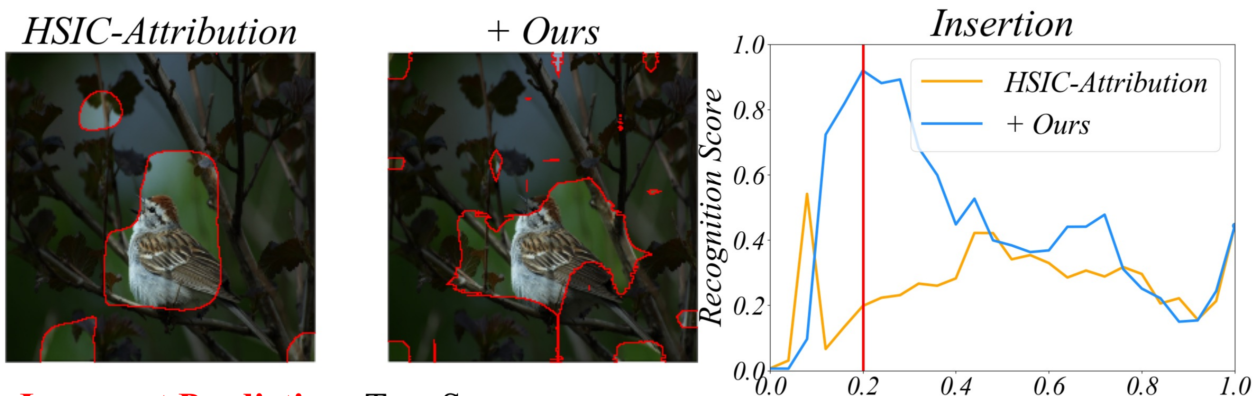


Incorrect Prediction: Hooded Oriole
Ground Truth: Orchard Oriole

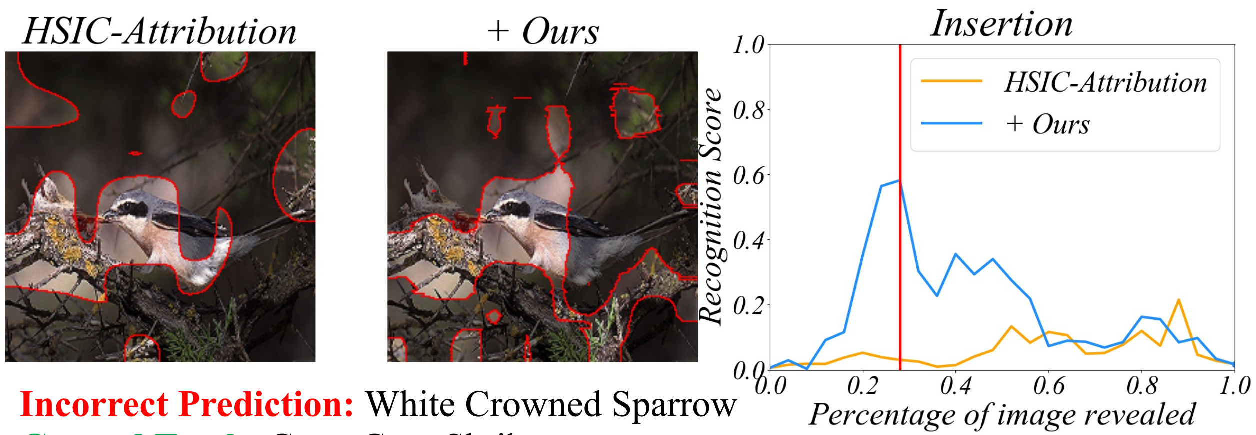
Debugging Model Prediction Errors



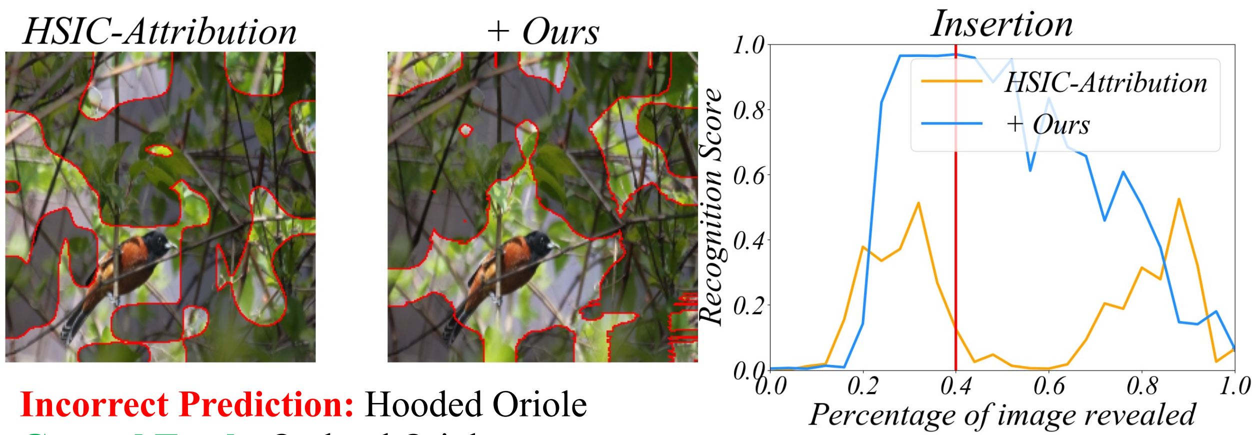
Incorrect Prediction: Tree Sparrow
Ground Truth: Lazuli Bunting



Incorrect Prediction: Tree Sparrow
Ground Truth: Chipping Sparrow



Incorrect Prediction: White Crowned Sparrow
Ground Truth: Great Grey Shrike



Incorrect Prediction: Hooded Oriole
Ground Truth: Orchard Oriole

Dark regions are the cause of model prediction errors

Debugging Model Prediction Errors

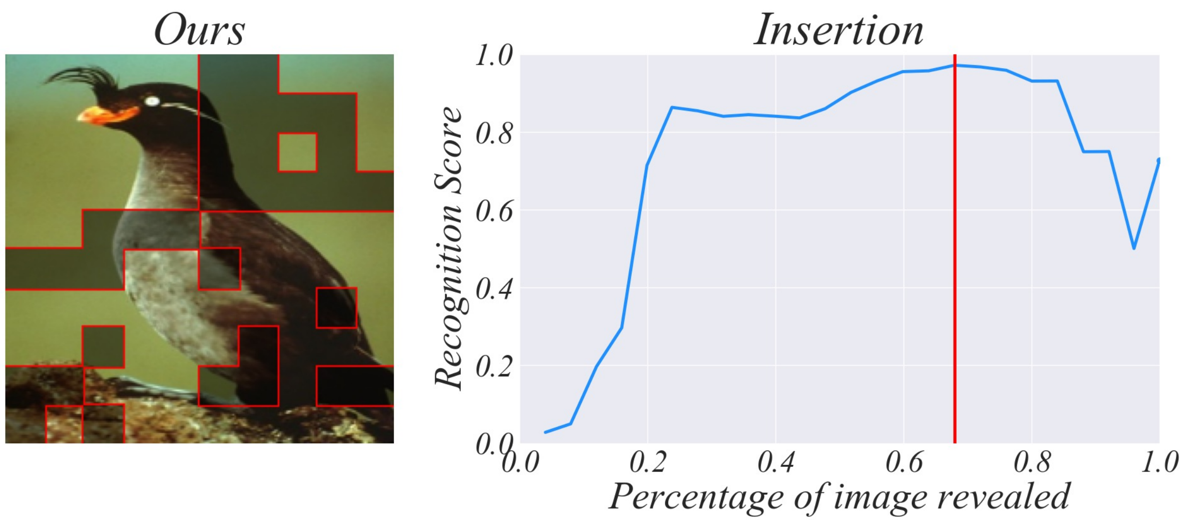
Table 2: Evaluation of discovering the cause of incorrect predictions.

Method	Average highest confidence (↑) (0-25%)	(0-50%)	(0-75%)	(0-100%)	Insertion (↑)
Grad-CAM++ (Chattopadhay et al., 2018)	0.1988	0.2447	0.2544	0.2647	0.1094
Grad-CAM++ (w/ ours)	**0.2424**	**0.3575**	**0.3934**	**0.4193**	**0.1672**
Score-CAM (Wang et al., 2020)	0.1896	0.2323	0.2449	0.2510	0.1073
Score-CAM (w/ ours)	**0.2491**	**0.3395**	**0.3796**	**0.4082**	**0.1622**
HSIC-Attribution (Novello et al., 2022)	0.1709	0.2091	0.2250	0.2493	0.1446
HSIC-Attribution (w/ ours)	**0.2430**	**0.3519**	**0.3984**	**0.4513**	**0.1772**

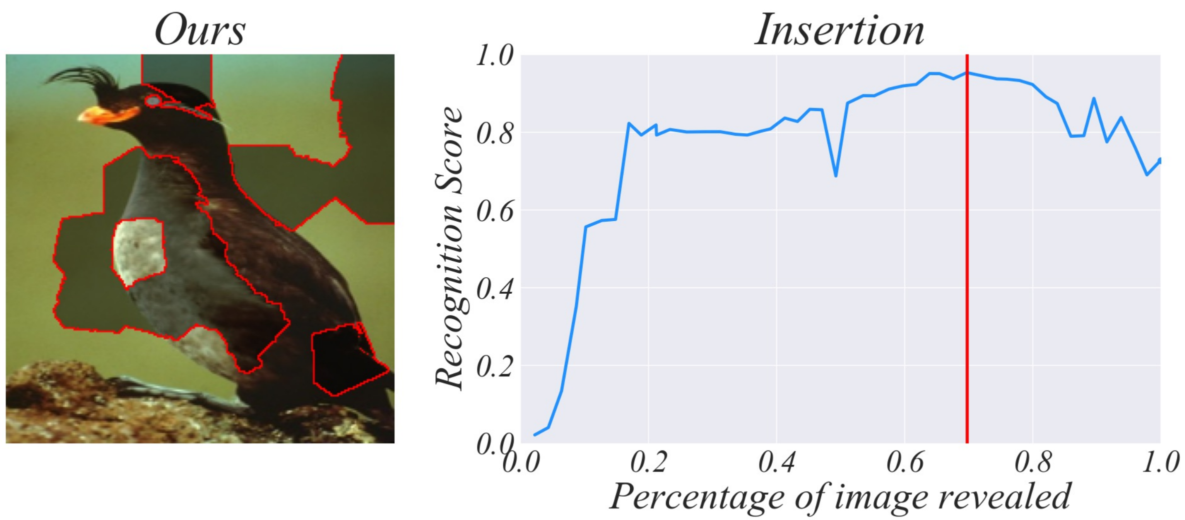
Average highest confidence: *67.3%* improvement

Insertion: *40.8%* improvement

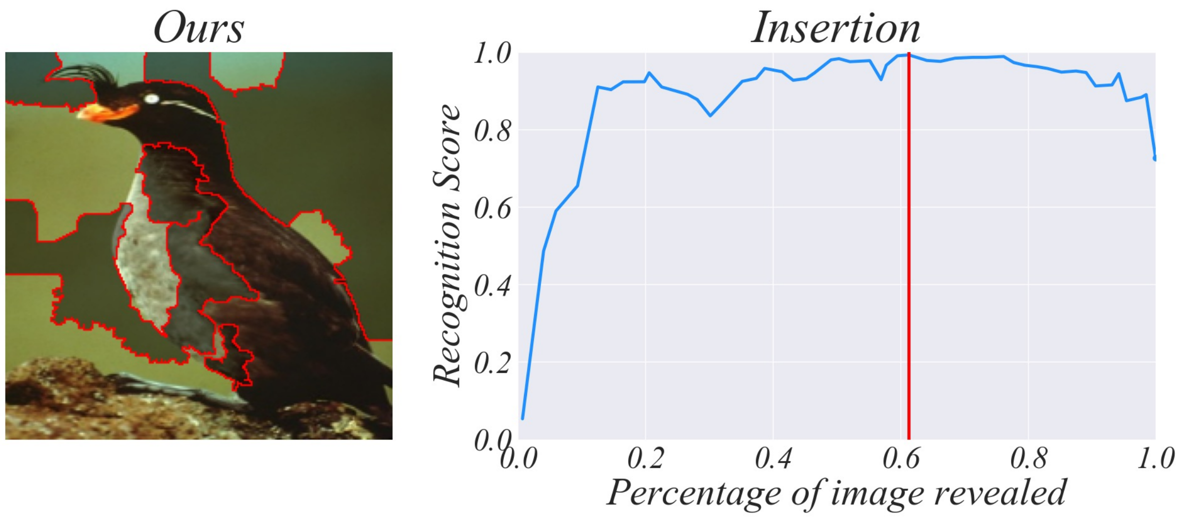
Extensions: division methods



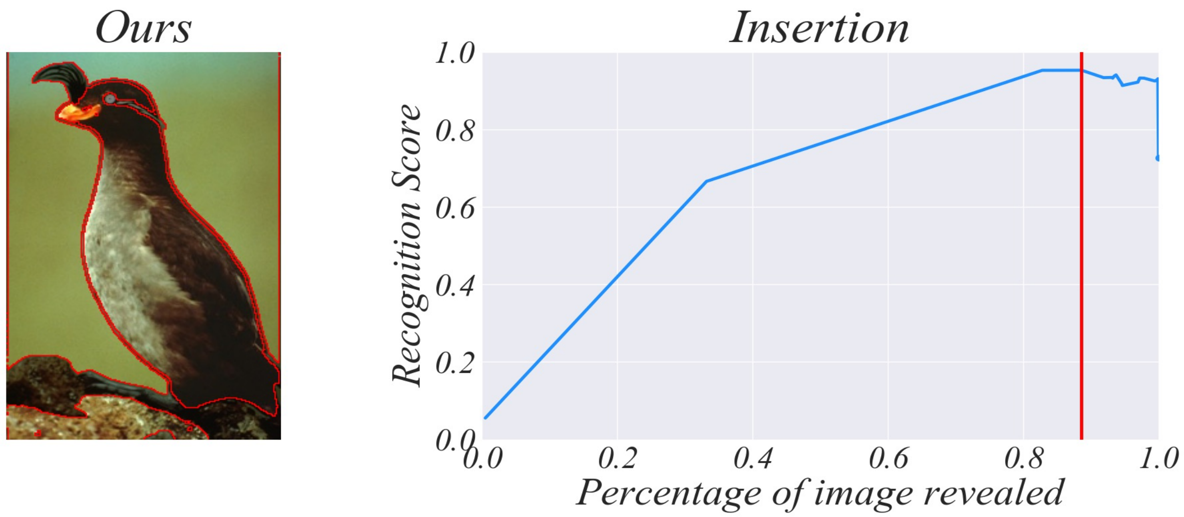
Prior Saliency Map (*This paper*), Insertion AUC: 0.7236



SLICO, Insertion AUC: 0.7604

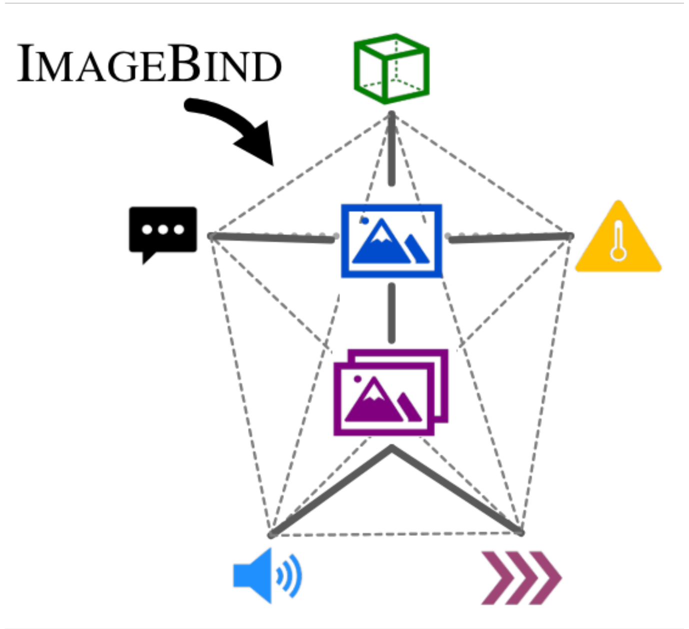


SEED, Insertion AUC: 0.8862

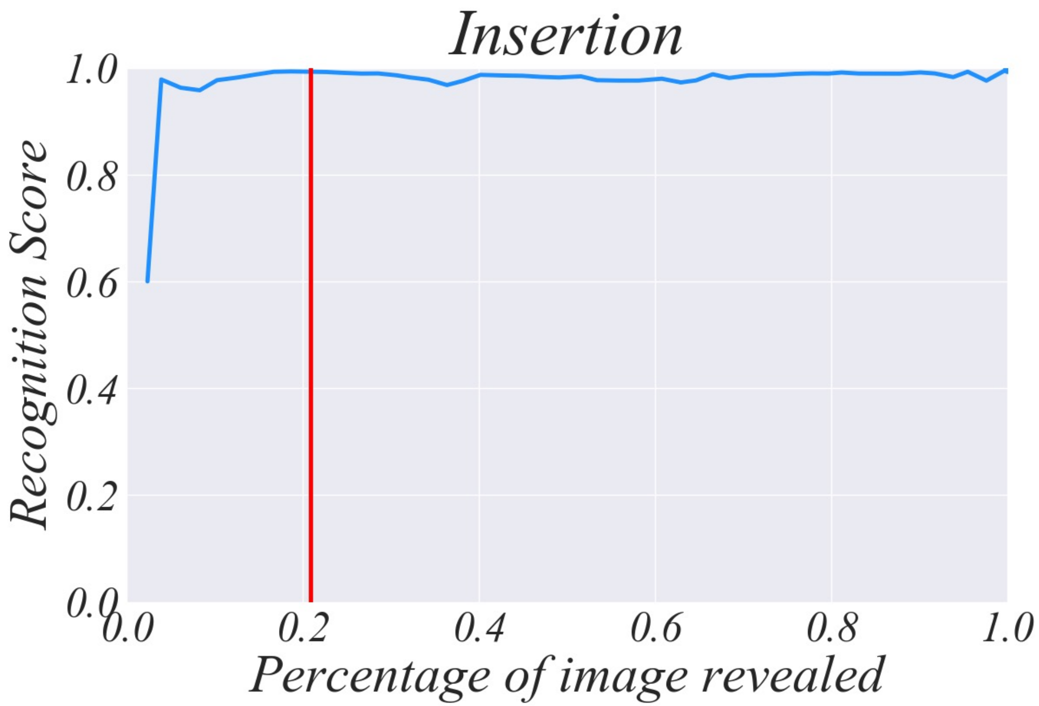


Segment Anything, Insertion AUC: 0.6803

Extensions: explaining multimodal foundation model



ImageBind is a Transformer-based multimodal model that can generate joint embeddings across seven modalities



Easy to scale to large model.

Summary

- A new perspective on image attribution: submodular subset selection
- A general attribution method for image classification problems that can be easily scaled to large models
- Can effectively discover potential regions that cause model's wrong prediction

Thank you so much for listening!

Poster: Hall B #219

Speaker: Ruoyu Chen
University of Chinese Academy of Sciences

R. Chen's Homepage　　WeChat　　Paper　　Code