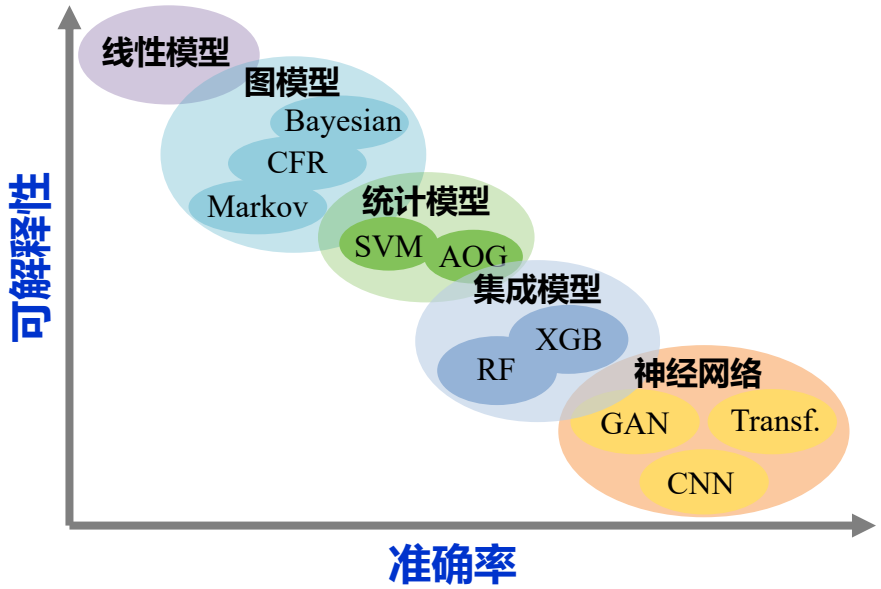
LESS IS MORE: FEWER INTERPRETABLE REGION VIA SUBMODULAR SUBSET SELECTION

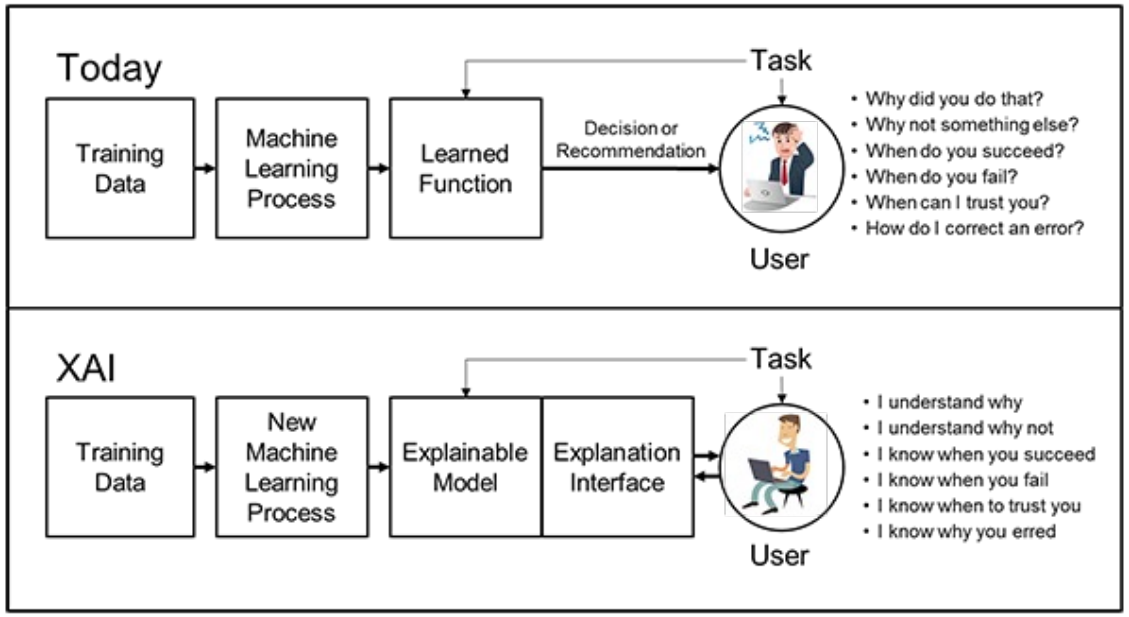
Ruoyu Chen[1,2], **Hua Zhang**[1,2,*], **Siyuan Liang**[3], **Jingzhi Li**[1,2], **Xiaochun Cao**[4,*]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
{chenruoyu,zhanghua,lijingzhi}@iie.ac.cn
[3]School of Computing, National University of Singapore, 119077, Singapore
pandaliang521@gmail.com
[4]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
caoxiaochun@mail.sysu.edu.cn

Accepted to ICLR 2024 **Oral (85/7262)**
Paper: https://arxiv.org/abs/2402.09164
Code: https://github.com/RuoyuChen10/SMDL-Attribution

研究背景

可解释AI





ML的巨大成功使AI的能力爆炸式增长，但其有效性将受到机器**无法向人类用户解释其决策和行动**的限制。**XAI**对于用户理解、适当信任和有效管理**新一代人工智能**至关重要。

可解释的人工智能 (XAI) 计划[1]：

- 产生更可解释的模型，同时保持高水平的学习性能（预测准确性）；
- 使人类用户能够理解、适当信任并有效管理新一代人工智能合作伙伴。

[1] Explainable Artificial Intelligence, https://www.darpa.mil/program/explainable-artificial-intelligence

相关术语

可解释AI

Interpretation

- 模型背后实际的**运行机理**；
- 准确将模型的原因与结果联系起来；
- 确定模型实际学习了什么；
- 在一定条件下是正确的。

Explanation

- 以**人类可理解**的方式表示决策过程或者结果；
- 关联各种反馈的模态，以及控制语义表达程度；
- 不一定是正确的。

Ante-hoc (拉丁语)

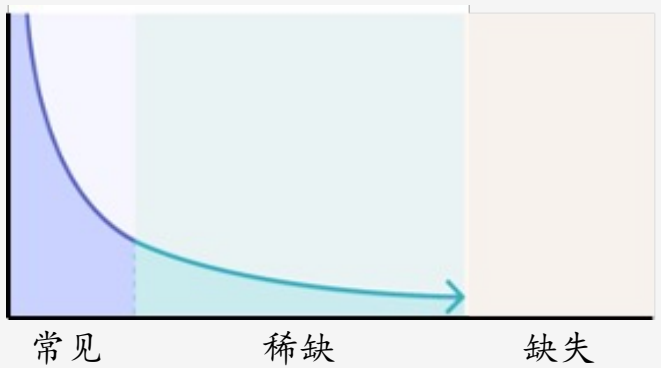
- 直接解释**白盒模型**；
- 在模型的决策过程中已产生可解释。

Post-hoc (拉丁语)

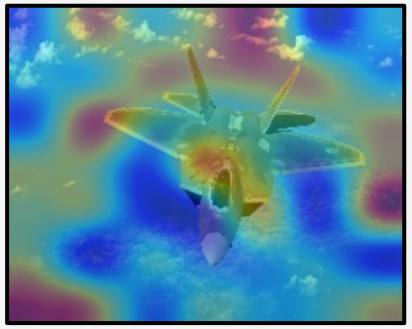
- 解释一个预训练模型或其决策的结果；
- 在模型做完决策后提供的解释。

现实挑战

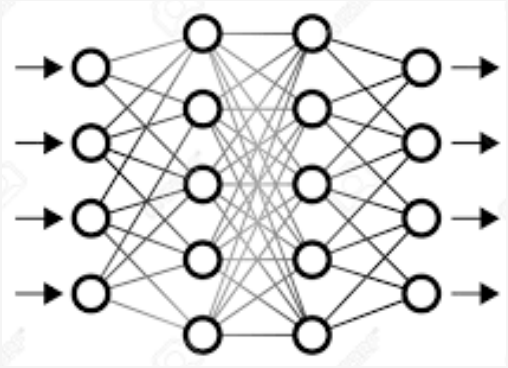
为什么AI模型仍存在错误？



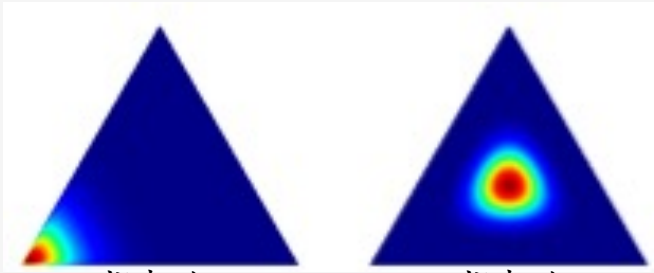
数据分布不全面



监督信息少



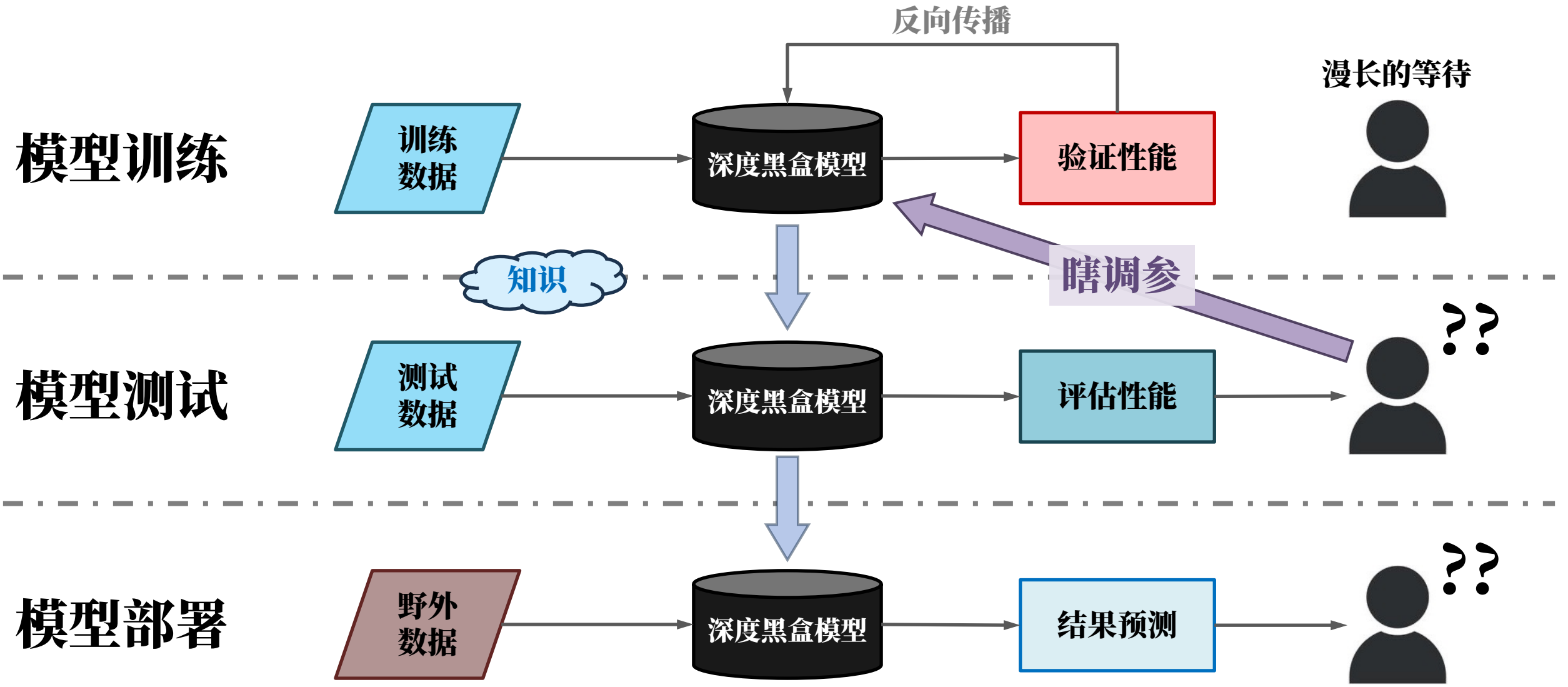
模型自身的缺陷



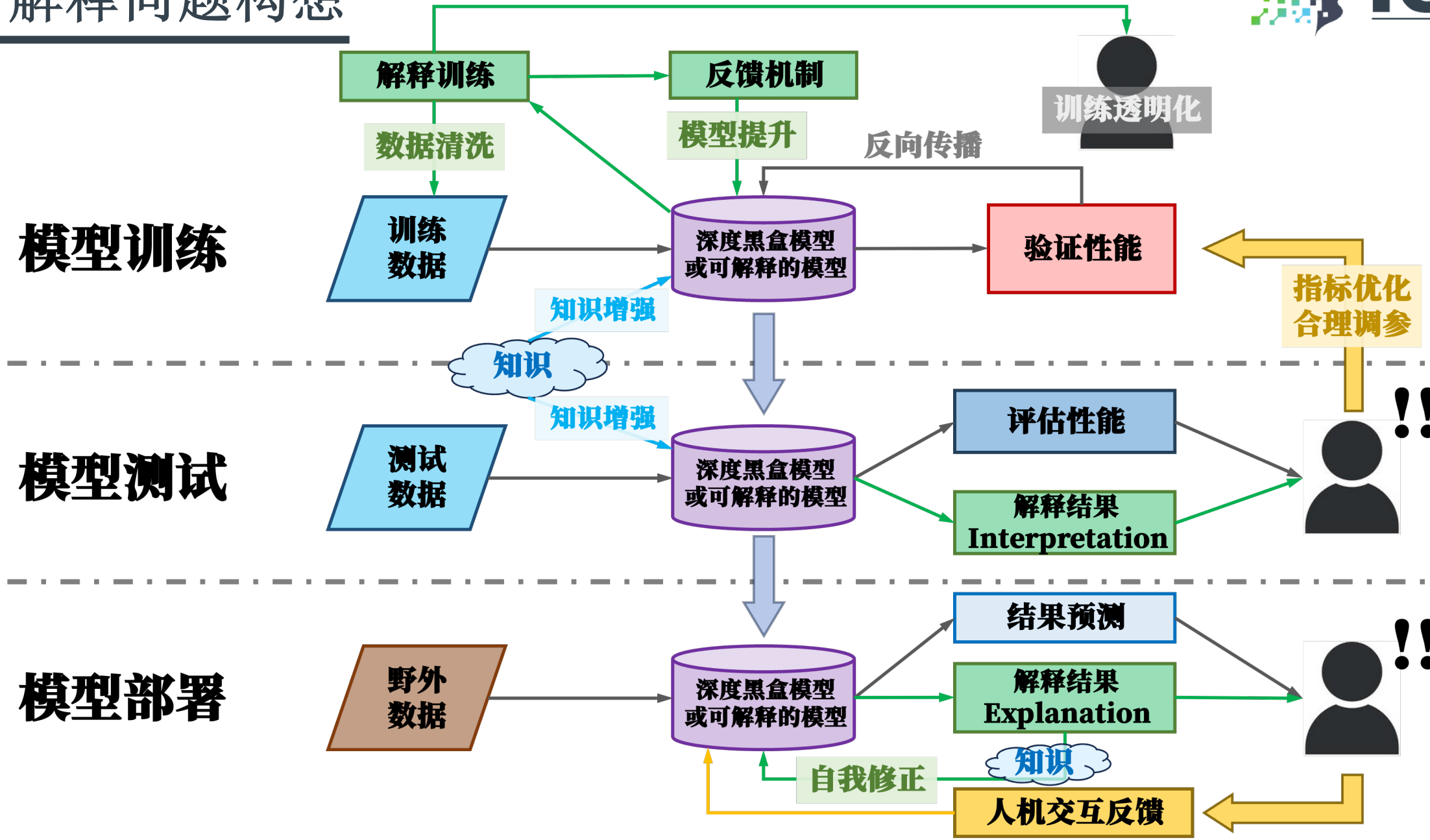
评价指标缺陷

所以我们需要可解释性！

可解释问题构想

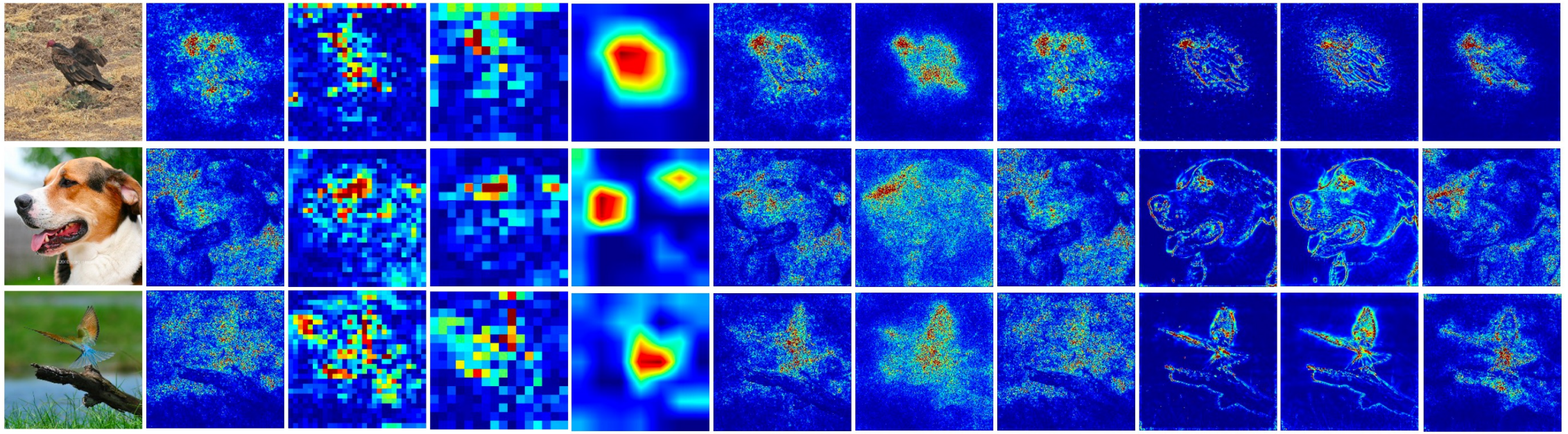

反向传播
漫长的等待
模型训练
训练数据
深度黑盒模型
验证性能
知识
瞎调参
模型测试
测试数据
深度黑盒模型
评估性能
??
模型部署
野外数据
深度黑盒模型
结果预测
??

可解释问题构想

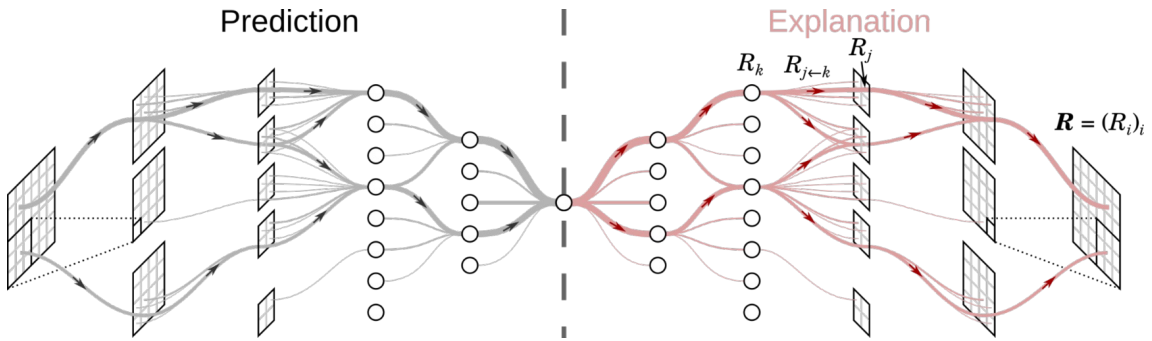

解释训练
反馈机制
训练透明化
数据清洗
模型提升
反向传播
模型训练
训练
数据
深度黑盒模型
或可解释的模型
验证性能
知识增强
知识
指标优化
合理调参
知识增强
模型测试
测试
数据
深度黑盒模型
或可解释的模型
评估性能
解释结果
Interpretation
!!
结果预测
模型部署
野外
数据
深度黑盒模型
或可解释的模型
解释结果
Explanation
!!
知识
自我修正
人机交互反馈

1. 研究背景

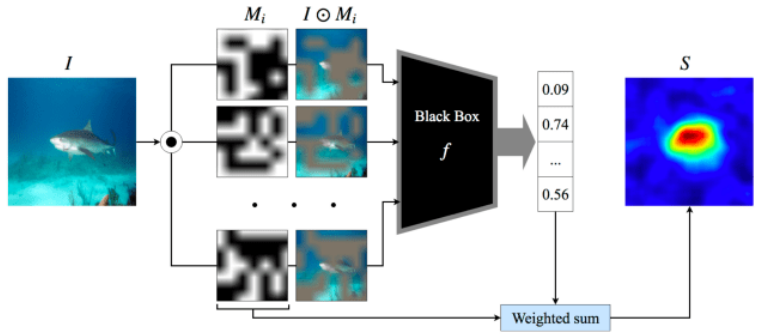
基于归因的方法



基于模型内部机理（白盒）



基于扰动（黑盒）



2. 问题挑战

问题：

- 现有的归因方法会产生*不准确的小区域*，从而误导正确归因的方向；
- 对于*预测错误*的样本，模型无法产生良好的归因结果。

解决方案：

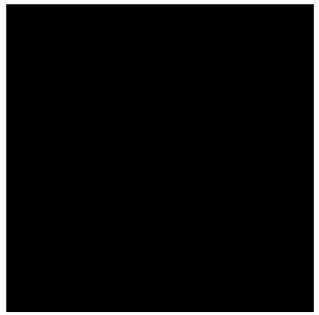
- 将归因问题重新表述为*子模子集选择*问题；
- 构建了一种新颖的*子模机制*。从模型的预测置信度、区域语义的有效性、语义的一致性和区域的集体效应四个方面来挖掘哪些区域促进了可解释性。这可以*限制对错误类响应区域的搜索*。

Input Image **I**

Attribution Map $\mathcal{A}$

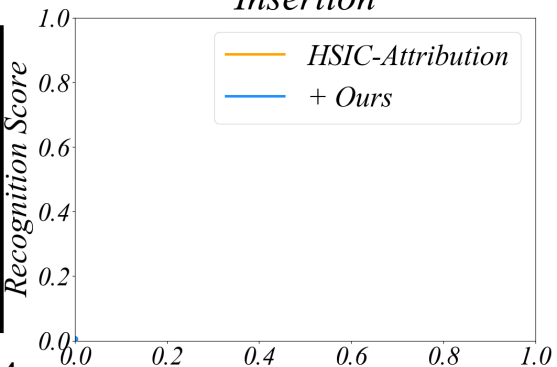
Sample correctly predicted by the model

HSIC-Attribution



+ Ours

Insertion

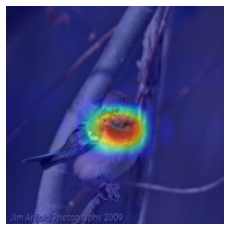


Original prediction confidence: 0.7284

Confidence in the region we discovered: **0.9543**

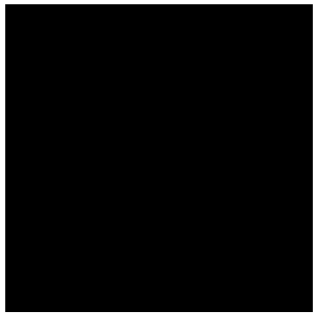
Input Image **I**

Attribution Map $\mathcal{A}$



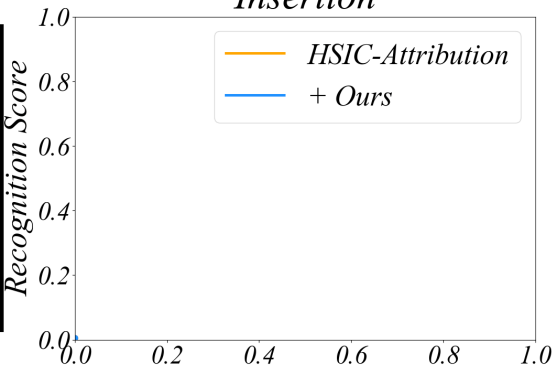
Sample incorrectly predicted by the model

HSIC-Attribution



+ Ours

Insertion



Incorrect Prediction: Tree Sparrow

Ground Truth: Lazuli Bunting

3. Preliminaries

定义 3.1 (Submodular function)

对于任意子集$S_a \subseteq S_b \subseteq V$，给定一个元素$\alpha$，其中$\alpha = V \backslash S_b$。当集合方程$\mathcal{F}$满足单调非负性

$$\mathcal{F}(S_b \cup \{\alpha\}) - \mathcal{F}(S_b) \geq 0,$$

以及：

$$\mathcal{F}(S_a \cup \{\alpha\}) - \mathcal{F}(S_a) \geq \mathcal{F}(S_b \cup \{\alpha\}) - \mathcal{F}(S_b)$$

时，$\mathcal{F}$为一个子模方程（submodular function）。

问题表述（子模子集选择理论）

给定一个集合V，一个子模方程$\mathcal{F}(\cdot)$，给定需要搜索的元素数量k，我们的目标是发现一个子集$S \subseteq V$，使子模方程数值最大化：

$$\max_{S \subseteq V, |S| < k} \mathcal{F}(S).$$

4. 提出方法

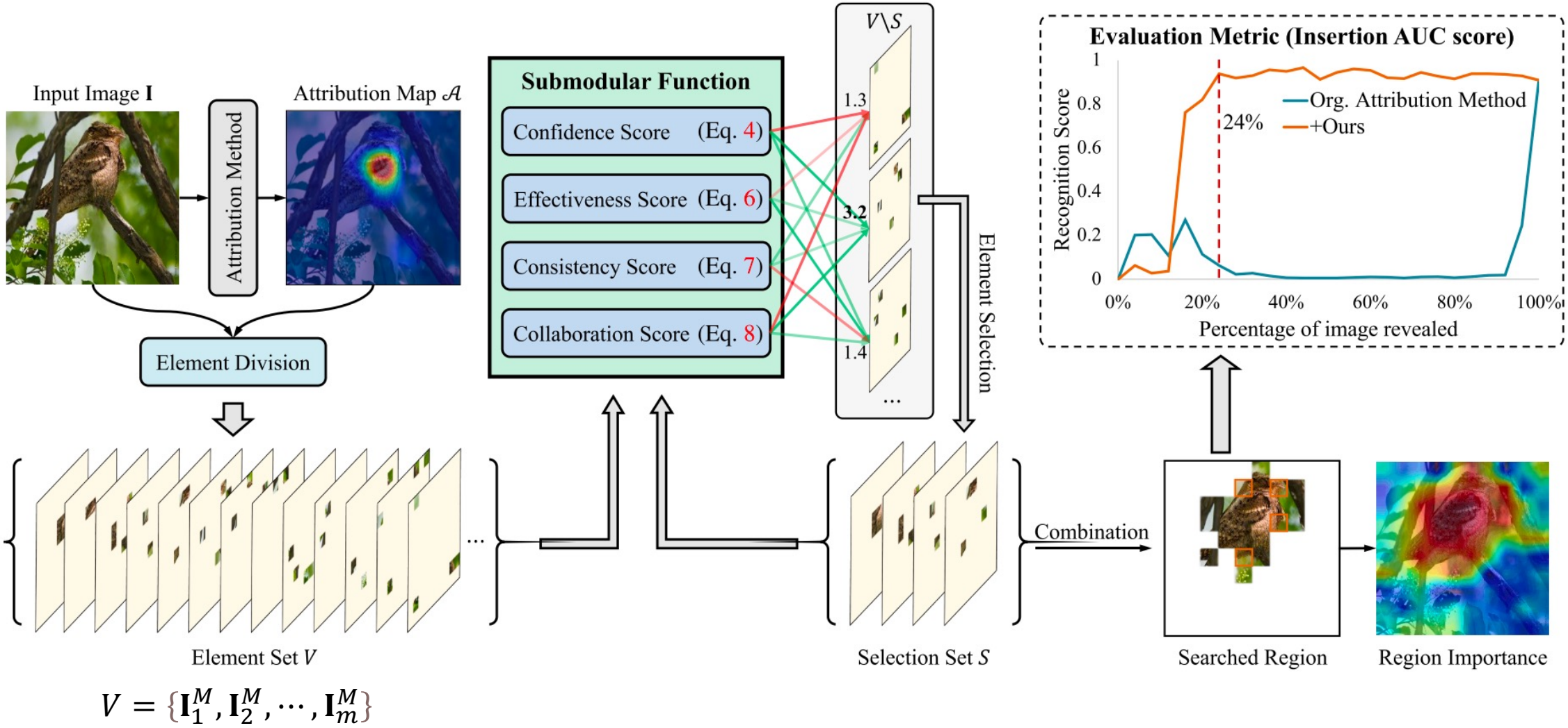


Figure 2: The framework of the proposed method.

4. 提出方法

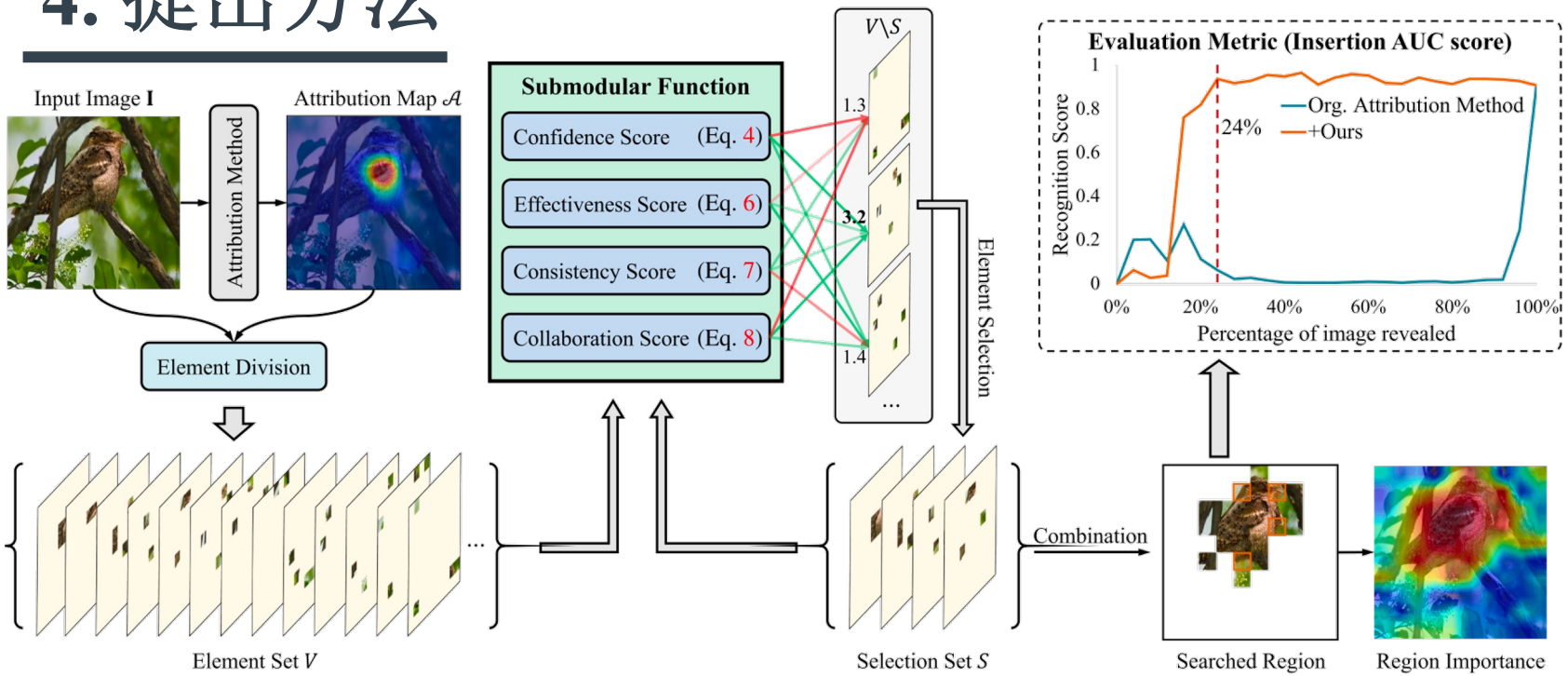


Figure 2: The framework of the proposed method.

引理 1

对于任意子集$S_a \subseteq S_b \subseteq V$，给定一个元素$\alpha$，其中$\alpha = V\backslash S_b$。方程$\mathcal{F}$满足子模方程性质：

$$\mathcal{F}(S_a \cup \{\alpha\}) - \mathcal{F}(S_a) \geq \mathcal{F}(S_b \cup \{\alpha\}) - \mathcal{F}(S_b)$$

引理 2

对于任意子集$S \subseteq V$，给定一个元素α，其中$\alpha = V\backslash S$。当集合方程$\mathcal{F}$满足单调非负性

$$\mathcal{F}(S \cup \{\alpha\}) - \mathcal{F}(S) \geq 0$$

设计子模方程 $\mathcal{F}(\cdot)$

- 置信度分数 (Eq. 4):

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

- 有效性分数 (Eq. 6):

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

- 一致性分数 (Eq. 7):

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

- 协作分数 (Eq. 8):

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

$$\mathcal{F}(S) = \lambda_1 s_{\text{conf.}}\left(\sum_{\mathbf{I}^M \in S} \mathbf{I}^M\right) + \lambda_2 s_{\text{eff.}}(S) + \lambda_3 s_{\text{cons.}}(S, \boldsymbol{f}_S) + \lambda_4 s_{\text{colla.}}(S, \mathbf{I}, \boldsymbol{f}_S)$$

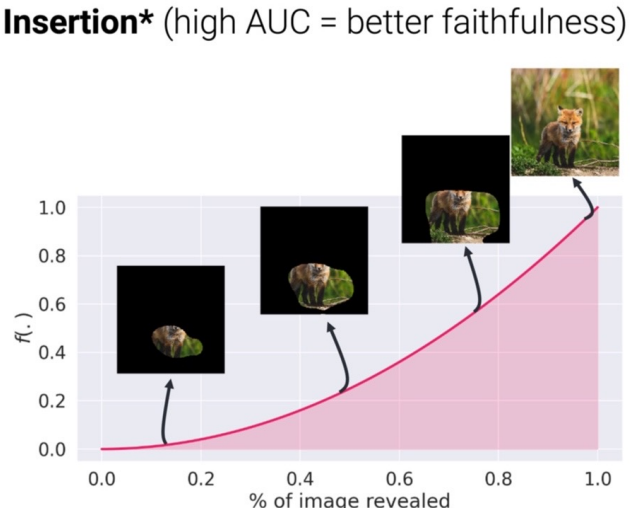
5. 实验验证

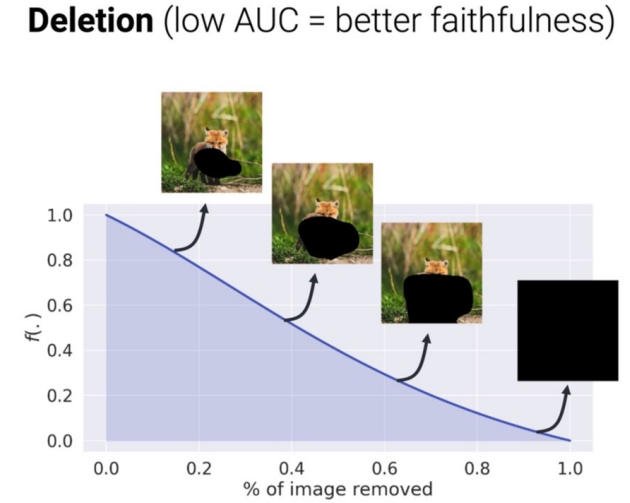
数据集：具有挑战性的人脸数据集VGG-Face2（8631个ID），Celeb-A（10177个ID）。细粒度数据集CUB-200-2011数据集（200个鸟类）。

评价指标：使用定量验证可解释结果Fidelity的指标Insertion AUC score和Deletion AUC score。

验证网络：VGGNet-19，ResNet-101，MobileNetV2和EfficientNetV2-M。

归因方法：Saliency，Grad-CAM，Grad-CAM++，Score-CAM，LIME，Kernel Shap，RISE和HSIC-Attribution（*NeurIPS 2022，SOTA方法*）。





5. 实验验证

1. 在网络正确预测样本上归因效果

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**
	8.4%	1.1%	1.0%	0.2%	5.3%	6.1%

5. 实验验证

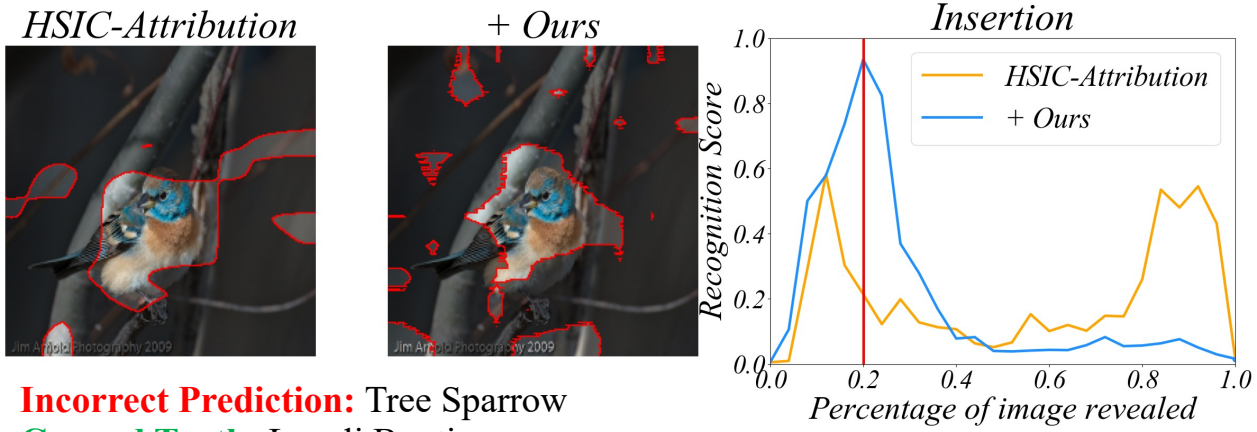
2. 在网络错误预测样本上归因效果

Table 6: Evaluation on discovering the cause of incorrect predictions for different network backbones.

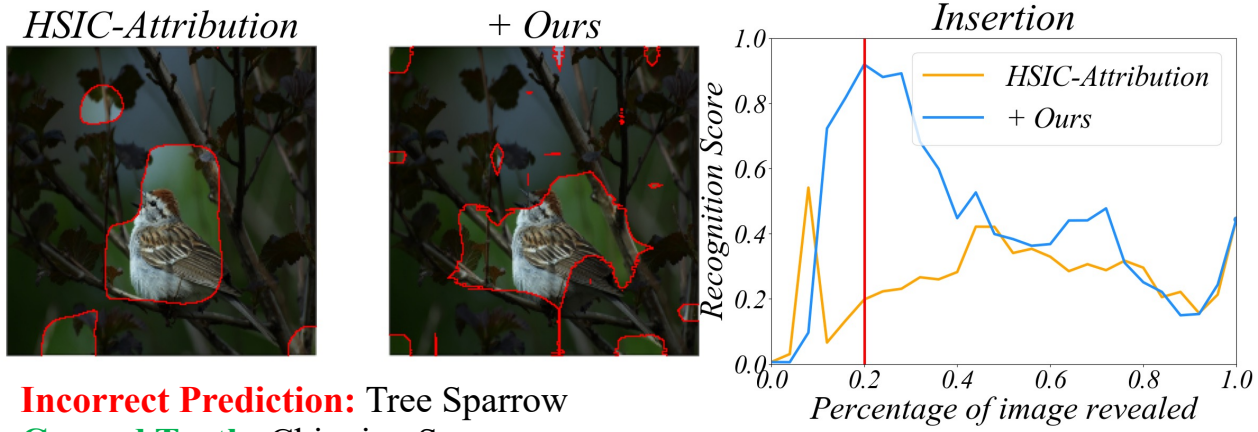
Backbone	Method	Average highest confidence (↑) (0-25%)	(0-50%)	(0-75%)	(0-100%)	Insertion (↑)
VGGNet-19 (Simonyan & Zisserman, 2015)	Grad-CAM++ (Chattopadhay et al., 2018)	0.1323	0.2130	0.2427	0.2925	0.1211
	Grad-CAM++ (w/ ours)	**0.1595**	**0.2615**	**0.3521**	**0.4263**	**0.1304**
	Score-CAM (Wang et al., 2020)	0.1349	0.2125	0.2583	0.3058	0.1057
	Score-CAM (w/ ours)	**0.1649**	**0.2624**	**0.3452**	**0.4224**	**0.1186**
	HSIC-Attribution (Novello et al., 2022)	0.1456	0.1743	0.1906	0.2483	0.1297
	HSIC-Attribution (w/ ours)	**0.1745**	**0.2716**	**0.3477**	**0.4226**	**0.1365**
ResNet-101 (He et al., 2016)	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**
MobileNetV2 (Sandler et al., 2018)	Grad-CAM++ (Chattopadhay et al., 2018)	0.1584	0.2820	0.3223	0.3462	0.1284
	Grad-CAM++ (w/ ours)	**0.1680**	**0.3565**	**0.4615**	**0.5076**	**0.1759**
	Score-CAM (Wang et al., 2020)	0.1574	0.2456	0.2948	0.3141	0.1195
	Score-CAM (w/ ours)	**0.1631**	**0.3403**	**0.4283**	**0.4893**	**0.1667**
	HSIC-Attribution (Novello et al., 2022)	0.1648	0.2190	0.2415	0.2914	0.1635
	HSIC-Attribution (w/ ours)	**0.2460**	**0.4142**	**0.4913**	**0.5367**	**0.1922**
EfficientNetV2-M (Tan & Le, 2021)	Grad-CAM++ (Chattopadhay et al., 2018)	0.2338	0.2549	0.2598	0.2659	0.1605
	Grad-CAM++ (w/ ours)	**0.2502**	**0.3038**	**0.3146**	**0.3214**	**0.1795**
	Score-CAM (Wang et al., 2020)	0.2126	0.2327	0.2375	0.2403	0.1572
	Score-CAM (w/ ours)	**0.2442**	**0.2900**	**0.3029**	**0.3115**	**0.1745**
	HSIC-Attribution (Novello et al., 2022)	0.2418	0.2561	0.2615	0.2679	0.1611
	HSIC-Attribution (w/ ours)	**0.2616**	**0.3117**	**0.3235**	**0.3306**	**0.1748**

5. 实验验证

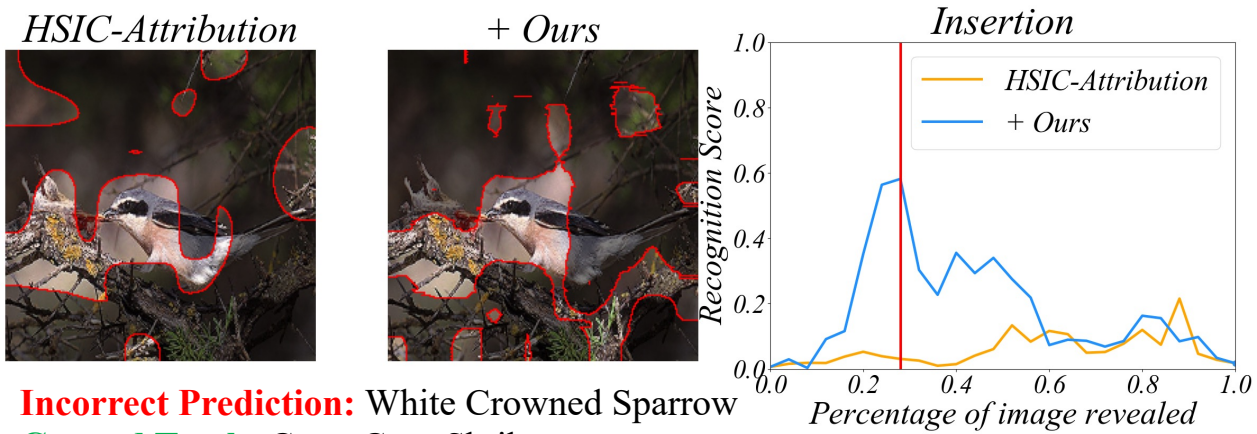
2. 在网络错误预测样本上归因效果



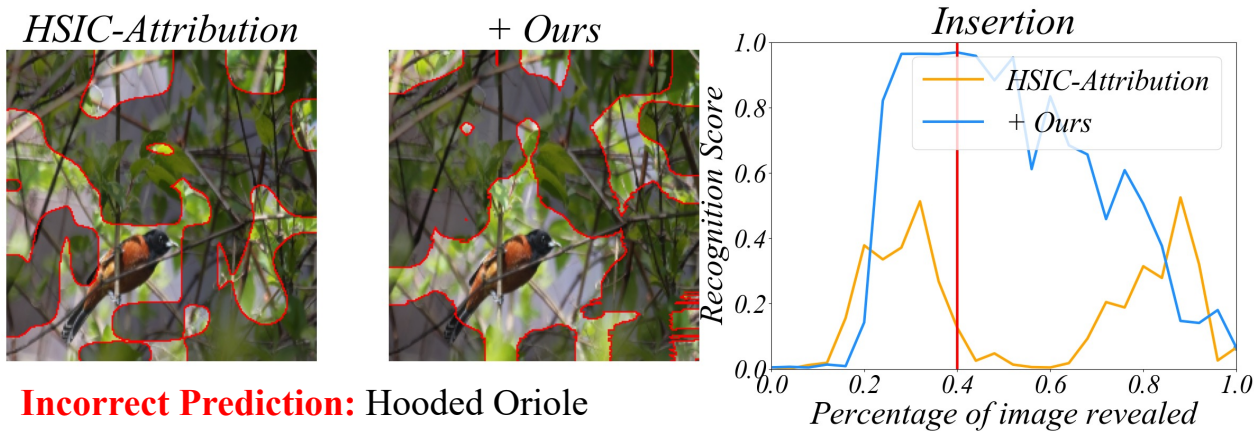
Incorrect Prediction: Tree Sparrow
Ground Truth: Lazuli Bunting



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

5. 实验验证

3. 消融实验

子模函数的有效性：

Table 3: Ablation study on components of different score functions of submodular function on the Celeb-A, and CUB-200-2011 validation sets.

Submodular Function				Celeb-A		CUB-200-2011	
Conf. Score (Equation 4)	Eff. Score (Equation 6)	Cons. Score (Equation 7)	Colla. Score (Equation 8)	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)
✓	✗	✗	✗	0.3161	0.1795	0.3850	0.3455
✗	✓	✗	✗	0.1211	0.5615	0.0835	0.6383
✗	✗	✓	✗	0.2849	0.2291	0.1019	0.6905
✗	✗	✗	✓	0.1591	0.3053	0.0771	0.5409
✓	✓	✗	✗	0.1075	0.5714	0.0865	0.6624
✗	✓	✓	✗	0.1082	0.5692	0.0750	0.7111
✗	✗	✓	✓	0.1558	0.3617	0.0641	0.7181
✗	✓	✓	✓	0.1074	0.5735	0.0632	0.7169
✓	✗	✓	✓	0.1993	0.2616	0.0623	0.7227
✓	✓	✗	✓	0.1067	0.5712	0.0651	0.6753
✓	✓	✓	✗	0.1088	0.5750	0.0811	0.7090
✓	✓	✓	✓	**0.1054**	**0.5752**	**0.0613**	**0.7262**

先验显著图划分的有效性：

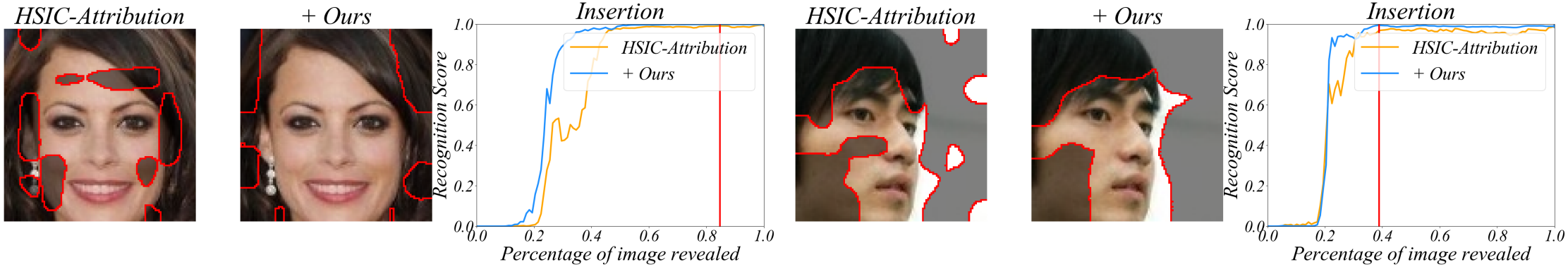
Table 4: Impact on whether to use a priori attribution map.

Method	Divided set size m	Celeb-A		CUB-200-2011	
		Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)
Patch 7×7	49	0.1493	0.5642	0.1061	0.6903
Patch 10×10	100	0.1365	0.5459	0.1024	0.6159
Patch 14×14	196	0.1284	0.5562	0.0853	0.5805
+HSIC-Attribution	25	**0.1054**	**0.5752**	**0.0613**	**0.7262**

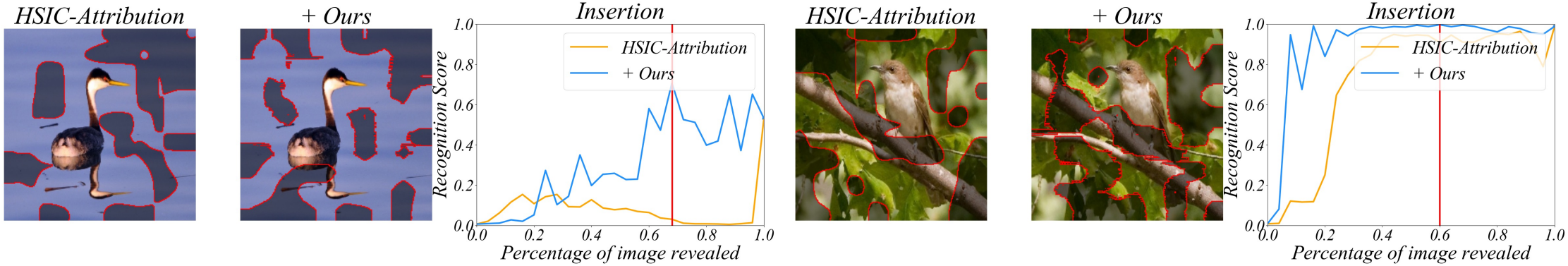
5. 实验验证

4. 可视化

Celeb-A

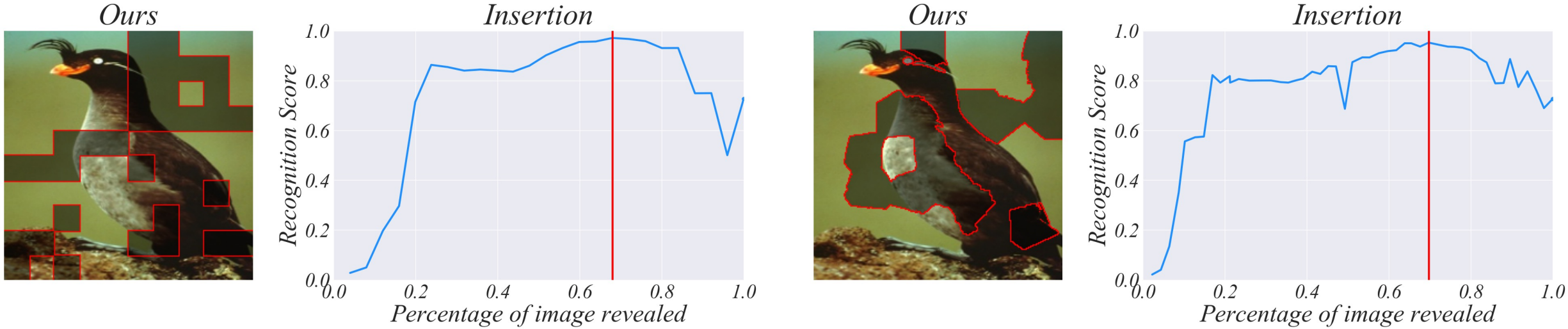


CUB-200-2011



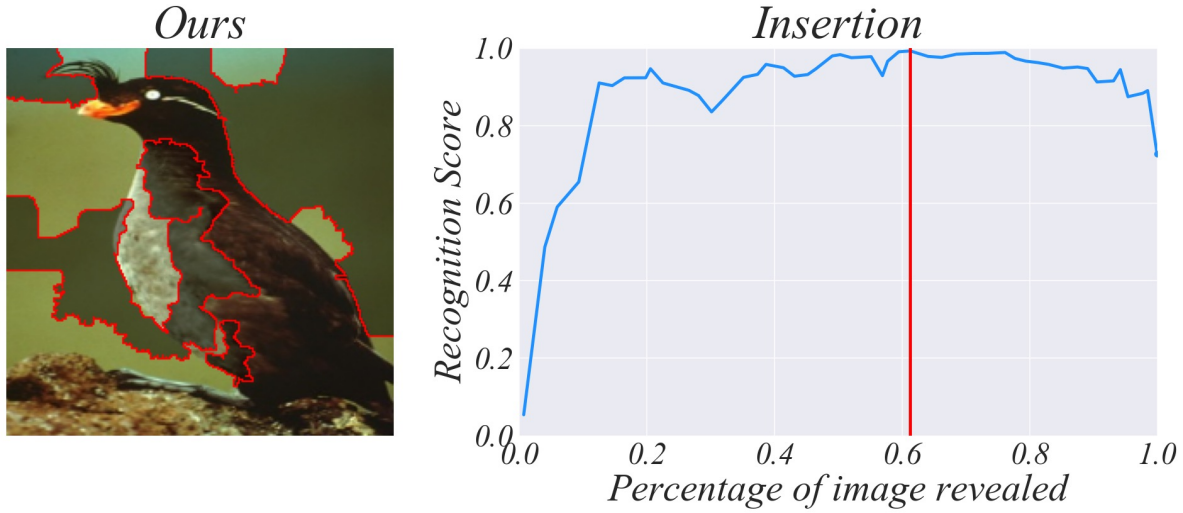
6. 未来展望

元素划分方法

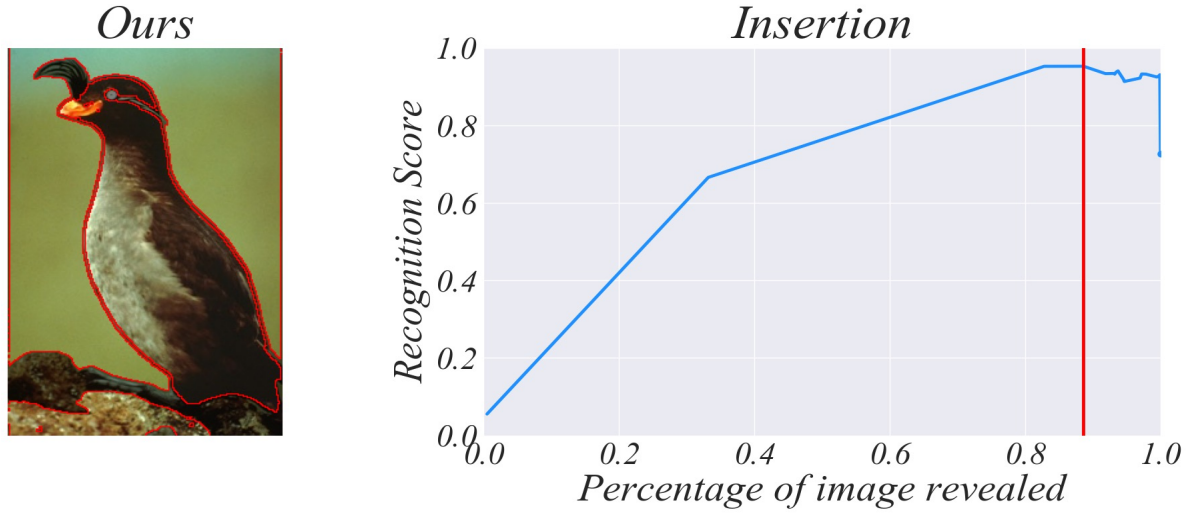


Prior Saliency Map, Insertion AUC: 0.7236

SLICO, Insertion AUC: 0.7604



SEED, Insertion AUC: 0.8862



Segment Anything, Insertion AUC: 0.6803

6. 未来展望

- 更多数据集，e.g., ImageNet，医学图像数据；
- 更多网络，e.g., Vision Transformer；
- 小模型到大模型，基础模型，多模态模型的验证；
- 更多元素划分方法，e.g., Segment Anything，超像素分割；
- 更多视觉任务， e.g., 目标检测；
- 等等。

仍有很多未知的可解释方法！

可解释性仍是一个有争议的话题！

更多的方法值得我们去探索发现！

欢迎大家加入可解释人工智能的研究！

谢谢各位聆听，
敬请批评指正！

Ruoyu Chen
China's Mainland

汇报人：陈若愚

Scan the QR code to add me as a friend.