Comprehensive Attribution: Inherently Explainable Vision Model with Feature Detector

Xianren Zhang[1], Dongwon Lee[1], and Suhang Wang[1]

The Pennsylvania State University, University Park, PA, USA
{xzz5508,dongwon,szw494}@psu.edu

Abstract. As deep vision models' popularity rapidly increases, there is a growing emphasis on explanations for model predictions. The inherently explainable attribution method aims to enhance the understanding of model behavior by identifying the important regions in images that significantly contribute to predictions. It is achieved by cooperatively training a selector (generating an attribution map to identify important features) and a predictor (making predictions using the identified features). Despite many advancements, existing methods suffer from the incompleteness problem, where discriminative features are masked out, and the interlocking problem, where the non-optimized selector initially selects noise, causing the predictor to fit on this noise and perpetuate the cycle. To address these problems, we introduce a new objective that discourages the presence of discriminative features in the masked-out regions thus enhancing the comprehensiveness of feature selection. A pretrained detector is introduced to detect discriminative features in the masked-out region. If the selector selects noise instead of discriminative features, the detector can observe and break the interlocking situation by penalizing the selector. Extensive experiments show that our model makes accurate predictions with higher accuracy than the regular blackbox model, and produces attribution maps with high feature coverage, localization ability, fidelity and robustness. Our code will be available at https://github.com/Zood123/COMET.

Keywords: Machine Learning · Explainability · Feature Attribtion

1 Introduction

Deep vision models have shown great performances across various Computer Vision tasks [10, 27, 28]. A critical requirement for deep vision models is the interpretability of their predictions. This is essential for experts and lay-users without the expertise to understand and trust the model. Moreover, interpretability can be imperative in high-stakes scenarios such as healthcare [15] and legal regulations [11, 16, 29]. Feature attribution, which identifies important input features that impact a model's prediction, can be understood by both experts and laypeople [3]. Feature attribution method typically generates attribution maps with the size of the input image, where each element of the map indicates the

importance of the corresponding pixel. There are two lines of attribution approaches, i.e., *post-hoc* and *inherently explainable* method [24, 33, 35, 48]. First, post-hoc methods aim to provide explanation for a pre-trained and fixed model. Generally, they aggregate information of the input image, activations [33, 48] and gradients [35] of the pre-trained model, and the output logits [23] to generate attribution maps. However, post-hoc approaches can not reduce the opacity of the pre-trained model where the decision process is still not revealed. Moreover, it has been shown that post-hoc attribution methods can be unfaithful, as the attribution map may not accurately reflect the true importance of the input features in making a prediction [19, 36]. To reveal how models use available features and guarantee the faithfulness of explanation, second, inherently explainable methods build models by jointly training a selector and a predictor [7, 13, 19]. Specifically, the selector first generates an attribution map that selects features and the predictor only adopts the selected features to make predictions. In this way, the feature attributions are more faithful because the prediction is only made by features selected by the attribution map.

However, existing inherently explainable methods have the *incompleteness problem* where not all discriminative features are highlighted. As shown in Fig. 1, existing explanation techniques tend to cover only parts of the target. For example, B-cos [5] focuses primarily on the beak of the duck, overlooking other parts. A recent study also shows that models tend to exploit shortcuts and only use 5% of input pixels for predictions [6]. In high-stakes scenarios, such as medical imaging, completeness is crucial for decision-making. The doctor may need the attribution map to know which part of the CT scan indicates the cancer result. The attribution map should contain all abnormal areas (e.g., tumors) to avoid misdiagnosis. The incompleteness problem arises from the current objective which only maximizes the prediction score. The objective encourages the selector to select features that are sufficient enough to support prediction (sufficiency) but not necessarily include all discriminative features (completeness).

Aside from the incompleteness problem, recent works [13, 19, 44] observe that joint training of the selector and the predictor tends to yield *degenerate cases* where the selector selects noise, resulting in low prediction accuracy. This issue arises from the interlocking problem of the selector and predictor [44]. Initially, the non-optimized selector would select noise and mask out discriminative features. The predictor then fits on the noise resulting in poor prediction accuracy. Conversely, the selector is updated to minimize the prediction loss by gradients that flow through the predictor. If the predictor has fitted on noise, the selector will be reinforced to preferentially choose the noise for low prediction loss. We also conduct an experiment on ImageNet9 to demonstrate this problem. The selector can select foreground or background. Similarly, the predictor can either fit on the foreground or background. Then, we evaluate the Cross-Entropy loss when the selector and the predictor use different strategies shown in Tab. 1. The strategy of selecting and fitting on the foreground (Foreground input, fit on F.) yields the globally minimal loss, which aligns with our goal for the model to focus on discriminative features. However, selecting and fitting on the background
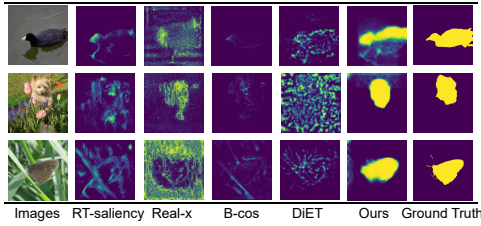
Fig. 1: Attribution maps from different explanation approaches.

Table 1: The Cross-Entropy (CE) loss when the predictor fits on the foreground(F.) or background(B.). The experiment is conducted on ImageNet-9 with ResNet18 as the predictor.

CE Loss	Predictor	
	fit on F.	fit on B.
Foreground input	**0.0027**	0.0298
Background input	0.0320	**0.0080**

(Background input, fit on B.) has lower loss than the other two strategies. This suggests that the model could easily fall into a sub-optimal interlocking state, selecting noise over discriminative features. The ground truth labels are the foreground objects and we assume that there are no spurious features in the background, which should consist solely of noise.

To address the aforementioned problems, we propose a novel inherently explainable model called COMprehensive fEature aTtribution (**COMET**). For the incompleteness problem, we enhance the comprehensiveness of feature selection with a new objective for masked-out regions. Alongside the existing objectives, we introduce an additional objective that discourages the presence of discriminative features in the masked-out regions. This means the selector is penalized if any discriminative features are not selected. This approach ensures a comprehensive and accurate attribution map, addressing the incompleteness problem effectively. For the interlocking problem, considering that the predictor can be misled and fit on the noise, we introduce a feature detector. This detector is designed to detect any discriminative features in the masked-out regions. If the selector mistakenly masks out discriminative features, concentrating instead on less important ones (noise), the predictor will fit on this noise, led astray by the selector. In such cases, the detector can identify these overlooked discriminative features and penalize the selector. The detector can break the interlocking between the selector and the predictor, avoiding degenerate cases. Additionally, our framework has several applications such as improving trustworthiness, helping model debugging [2, 26], and image generation [31]. Detailed discussions and examples are in the supplementary materials.

Our **main contributions** are as follows: (i) We introduce a novel inherently explainable model that improves feature selection comprehensiveness by discouraging the exclusion of discriminative features. Our method has more comprehensive attribution maps (Fig. 1); (ii) A feature detector that is pre-trained on full images to address the interlocking problem, allowing it to identify discriminative features in masked-out regions and break the interlocking cycle; (iii) Our extensive experiments demonstrate that our model outperforms various baselines in predictive accuracy and attribution map quality.

2   Related Work

Attribution methods explain deep vision models by identifying input regions that are important to the model's decision. Existing attribution methods can be categorized into Post-hoc methods and Inherently Explainable methods.

Post-hoc Methods. In post-hoc methods, the model is pre-trained and fixed. Post-hoc method can be generally divided into three categories. The first category is activation-based methods [18, 20, 25, 33, 48], which rely on the activation values of the internal neurons to explain the prediction. The first work [48] named Class Activation Mapping (CAM), averages the output of the final convolutional layer (activations) to generate the attribution map. Score-CAM [38] uses activations as masks and averages layer-wise class activations with classification scores. Ablation-CAM [25] generates attribution maps by ablating (removing) feature map units in the final convolutional layer and measuring the drop in class activation scores. The second category is backpropagation-based methods [21, 33–35, 41], which assumes that the importance of inputs is often measured by the gradients to inputs. However, gradients are found to be independent of the discriminative models and can not serve as valid explanations [36]. The third category is perturbation-based methods [9, 23, 30, 37], which identify the relevant features by removing (or maintaining) parts of the input features and summarizing changes in classification scores. RISE [23] generates attribution maps by randomly masking the input image and computing the weighted sum of these masks. To save inference time, a selector is trained to generate attribution maps. The selector is trained to preserve the classification score on the masked-in features of the image [9]. Recently, the selector is further modified to have class-wise output and better fitted into multi-label prediction scenario [37].

Inherently Explainable Methods. Inherently explainable methods train the selector and predictor jointly [7, 13, 19, 42]. Specifically, the selector generates attribution maps that select discriminative features. The selected features are then fed into a predictor to conduct predictions. Compared with post-hoc methods, they can generate more faithful attribution maps because the predictor only uses selected features to conduct predictions. However, as mentioned in the introduction, degenerate cases will happen because of the interlocking between the selector and the predictor [13, 19]. To prevent the selector mislead the predictor, REAL-X [19] trains the predictor model with randomly masked images. Then, it fixes the predictor and trains the selector by maximizing the classification score. RB-AEM [13] argues that random masks ignore the geometric prior where nearby pixels share similar semantic information. As a result, RB-AEM trains the predictor with random Radial Basis Function style masks. These methods can mitigate degenerate cases by isolating the predictor training process from the selector. However, there is a distribution shift where the features selected by the selector differ from randomly selected ones. The predictor which is trained on randomly masked images would have sub-optimum performance when tested on features selected by the selector. *Different from these approaches, our proposed approach can avoid degenerate cases while training the predictor and the selector jointly, maintaining high prediction performance.*
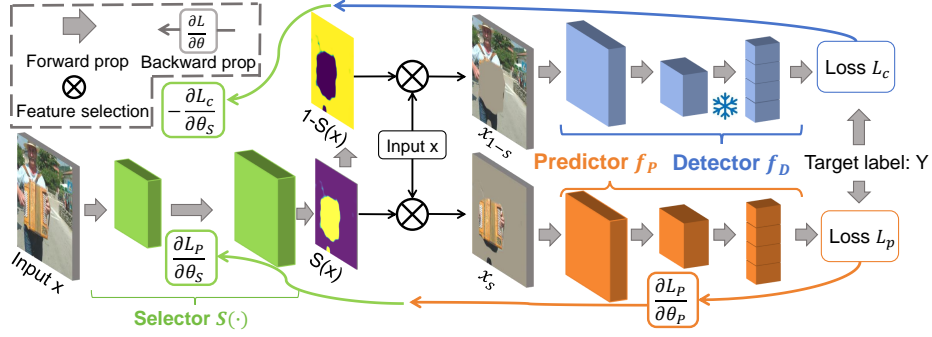
Fig. 2: The framework consists of a selector S, predictor f_P, and a pre-trained detector f_D. The selector first identifies discriminative features and generates the attribution map $S(x)$. The $S(x)$ selects features x_s, which are then fed into the predictor. Meanwhile, the remaining parts of the image, represented as x_{1-s}, are fed into the feature detector. The selector is optimized to assist the predictor in accurately predicting the label while confusing the detector by not leaving any discriminative features in x_{1-s}.

3 The Proposed Framework: COMET

We propose a novel framework, COMET, as illustrated in Fig. 2. Our framework comprises three modules: the selector S, the predictor f_P, and the detector f_C. The selector S identifies discriminative features and generates the attribution map, which is used to select important features for explainable prediction. These selected features are then fed into the predictor for label prediction. The remaining parts of the image are fed into a feature detector introduced to detect discriminative features in the masked-out regions. The detector is pre-trained on full images, enabling it to detect discriminative features without being influenced by the selector. The selector is optimized to assist the predictor in accurately predicting the label while confusing the detector by ensuring that masked-out regions do not contain discriminative information.

We focus on the image classification task. Let $x \in \mathbb{R}^{H \times W \times C}$ be an input image with H, W, and C being the height, width, and number of channels. Let $y \in [1, 2, 3, ...K]$ be the ground-truth label of x with K-being the number of classes. We represent the attribution map as $S(x) \in [0, 1]^{H \times W}$ where $S(\cdot)$ is the feature selector and each element of $S(x)$ represents the importance of the corresponding pixel. The attribution map is used as a mask to select discriminative features. Specifically, the masked-in features x_s and masked-out features x_{1-s} are represented as:

$$x_s = S(x) \odot x + (1 - S(x)) \odot q, \quad x_{1-s} = (1 - S(x)) \odot x + S(x) \odot q, \qquad (1)$$

where q is the matrix of the average value of pixels in the dataset. x_s is the masked-in features and x_{1-s} is the masked-out features.

3.1   The Predictor and Feature Detector

Since we want to build an inherently explainable model, different from the post-hoc explanation method, the predictor is trained with masked-in features so that it can conduct prediction tasks cooperatively with the selector. Specifically, we train the predictor with the following loss function:

$$\min_{P(\cdot)} \mathcal{L}_P(x_s) = H(p(y|x), f_P(x_s)), \tag{2}$$

where $H(\cdot, \cdot)$ is the cross-entropy and $p(y|x)$ is the ground-truth label. The predictor is optimal when correctly predicting the target with selected features.

Next, considering the interlocking between the selector and the predictor, our method introduces a feature detector as an independent check. Trained with full images, the detector remains unaffected by the selector's decisions. We train the detector with the following objective:

$$\min_{D(\cdot)} \mathcal{L}_D(x) = H(p(y|x), f_D(x)). \tag{3}$$

The detector is optimal when it can accurately predict the target label. Training on full images x allows the detector to observe features beyond those selected by the selector x_s (same for x_{1-s}). This is crucial when the selector's choices are sub-optimal or unreliable. If the selector masks out important features and the predictor fits on the noise, the predictor itself can not effectively detect the discriminative features in masked-out regions. At this time, the detector is needed to identify any discriminative features in the masked-out regions and break the interlocking between the selector and the predictor by penalizing the selector. We will show how the detector can break the interlocking between the selector and the predictor in the following section.

3.2   Optimizing the Feature Selector

With the predictor and the detector, we then introduce how to optimize the selector. The goal of the selector is to generate the attribution map. We first define what is an ideal attribution map. Then, we demonstrate how our training framework optimizes the selector to generate this ideal attribution map.

Definition 1. *(Ideal Attribution Map). For every input image and label (x, y), an optimal attribution map $S(x)$ should separate discriminative features (signal) and non-discriminative features (noise). There are two requirements for the ideal attribution map:*

- *Sufficiency: x_s contains sufficient discriminative features: $H(y|x_s) = H(y|x)$, where $H(y|x_s)$ is the conditional entropy of the label y given the selected discriminative features x_s. It requires the label information provided by the selected features x_s is equivalent to that provided by the entire image.*
- *Completeness: x_{1-s} only contains non-discriminative features: $H(y|x_{1-s}) = H(y)$. This means that the label information provided by the masked-out regions is no different from a random guess.*

We propose the attribution generation as a signal-noise decomposition process. A good attribution map should not only contain enough features for label prediction (sufficiency) but also try to include all discriminative features (completeness). With the definition of ideal attribution maps, we then introduce how our learning framework can optimize the selector to achieve the sufficiency and completeness requirements. We train the selector with the following objective:

$$\min_{S(\cdot)} \mathcal{L}_S(x) = \mathcal{L}_P(x_s) - a * \mathcal{L}_D(x_{1-s}), \tag{4}$$

where the first term is to help the predictor predict the label and the second term is to discourage the detector from predicting the label accurately for the masked-out regions.

Interlocking Mitigation. The second term is crucial for addressing the interlocking problem. When the selector selects noise, leading the predictor to fit on this noise, the system reaches a local optimum with a low $\mathcal{L}_P(x_s)$. In this scenario, the detector, which is already trained with full images, can observe and identify the discriminative features that the selector has overlooked. The detector can approximate the label based on discriminative features in masked-out regions, even though it has not been explicitly trained on masked images. As a result, $\mathcal{L}_D(x_{1-s})$ will be low, leading to a high selector loss $\mathcal{L}_S(x)$. This high loss signals the need for the selector to adjust its choices, encouraging it to escape the local optimum and select more discriminative features, thereby breaking the interlocking problem and aligning with the requirements of sufficiency and completeness. With certain assumptions, we then prove the optimal selector can satisfy these requirements.

Proposition 1. *(Optimal Selector). Assume that the predictor and the detector can effectively predict the labels of their input, i.e., $f_P(x_s) = p(y|x_s)$ and $f_D(x_{1-s}) = p(y|x_{1-s})$, then the optimal selector satisfies both the sufficiency and completeness requirements.*

Remark. The $p(y|x_s)$ and $p(y|x_{1-s})$ refer to the label given the masked-in (x_s) and masked-out (x_{1-s}) regions of the image. There are three situations. (i) Optimal Attribution Map: All discriminative features are selected for x_s and only noise is left for x_{1-s}. In this case, the label of masked-in regions is the same as the full image (i.e., $p(y|x_s) = p(y|x)$) and the label of the masked-out regions should be a uniform distribution (i.e., $p(y|x_{1-s}) = p(y)$). This situation means all discriminative features are selected and the attribution map is already ideal. (ii) Worst Attribution Map: Only noise is selected and discriminative features are all in masked-out regions. As discussed in the previous paragraph, the detector can approximate the label based on the discriminative features, resulting in low $\mathcal{L}_D(x_{1-s})$. This increases the selector loss $\mathcal{L}_S(x)$, which compels the selector to include discriminative features. Therefore, our loss function inherently discourages the attribution maps to only include noise, ensuring this scenario is rarely encountered. (iii) Partial Feature Selection: Discriminative features exist

both in masked-in and masked-out regions (e.g., A dog's head is included in x_s and its body is x_{1-s}). The labels for both masked-in and masked-out regions are the same as full images (i.e., $p(y|x_s) = p(y|x_{1-s}) = p(y|x)$). In this situation, $f_P(x_s) = p(y|x_s)$ holds when predictor is optimized with Eq. (2). As introduced in the previous section, the detector is trained with full images. With the understanding of the complete images, the detector can approximate the label based on discriminative features in masked-out regions (i.e., $f_D(x_{1-s}) = p(y|x_{1-s})$).

Proof. Since the predictor and detector can effectively predict the labels with their inputs:

$$\mathcal{L}_P(x_s) = H(p(y|x_s), f_P(x_s)) = H(y|x_s) \geq H(y|x), \qquad (5)$$

$$-\mathcal{L}_D(x_{1-s}) = -H(p(y|x_{1-s}), f_D(x_{1-s})) = -H(y|x_{1-s}) \geq -H(y). \qquad (6)$$

Thus, $\mathcal{L}_S \geq H(y|x) - a * H(y)$. The equality can be reached when $S(x) = S^*(x)$, where $S^*(x)$ is the ground-truth mask. $\mathcal{L}_S$ achieves equality if and only if both (Eq. (5)) and (Eq. (6)) reach their equalities. Thus, the sufficiency ($H(y|x_s) = H(y|x)$) and the completeness ($H(y|x_{1-s}) = H(y)$) requirements can be satisfied when the selector is optimized.

3.3   Final Objective Function of COMET

We demonstrate that our framework can generate sufficient and complete attribution maps while making predictions. In practice, a regularizer $R(S(x))$ is added to the selector loss: $\mathcal{L}_S + b * R(S(x))$ to control the sparsity of the attribution map and avoid the trivial example: $x_s = x$. The overall loss for our training framework:

$$\min_{S(\cdot), f_P(\cdot)} \mathcal{L}(x) = \mathcal{L}_P(x_s) - a * \mathcal{L}_D(x_{1-s}) + b * R(S(x)). \qquad (7)$$

Directly using the L1 norm as the regularizer is not optimal. Instead, we use a threshold t to control the sparsity of the mask: $R(S(x)) = max(\frac{1}{HW} \sum S(x) - t, 0)$. If over t fraction of the pixels are selected, the regularizer will penalize the model. The first term of the loss function is to predict the target label via the collaborative work of the selector and predictor. This term ensures that the attribution map contains enough information about the target (Sufficiency). The second objective term is to penalize the selector if the detector f_C can infer the target label with masked-out regions. This term discourages the presence of discriminative features in the masked-out regions (Completeness). An overall training algorithm is in the supplementary materials.

Discussion Recently, several other methods have been trying to mitigate the interlocking between the selector and the predictor. Instead of using the selected features x_s, they train the predictor with randomly masked images $x_{s'}$, either uniformly at random ($s' \sim$ Bernoulli(0.5)) as suggested by [19] or using a Radial Basis Function (RBF)-style masking ($s' \sim$ Bernoulli(RBF)) as proposed by [13]. Then, they fix the predictor while training the selector. These approaches mitigate the interlocking problem by isolating the training process of the predictor

from the selector. However, these mitigation approaches are sub-optimal because the predictor is trained with $x_{s'}$ (randomly selected features) which still differs from the distribution of x_s (selector selected features). The distribution shift would lead to low prediction accuracy and a less precise attribution map.

Learning with three modules has been employed in various tasks, such as domain adaptation [12, 46], learning sleep stages radio signals [47], and rationalization in text data [43]. These works use an adversarial framework to learn three modules, aiming to obtain domain-invariant representations or rationales of the input text. In contrast, our work focuses on improving the attribution maps of deep vision models. Additionally, unlike the adversarial frameworks in previous studies, our feature detector is pre-trained on full images and remains fixed during the training of the other two modules. Training the detector on masked-out regions would cause it to fit the noise in these areas, reducing its effectiveness in detecting discriminative features. By pre-training on complete images, the detector maintains its ability to see the entire picture without being misled by noise. To ensure the detector provides reliable signals from the start of the training process, it is pre-trained and kept fixed.

4 Empirical Validation

In this section, we conduct experiments to answer the following research questions: (RQ1) Can our proposed model conduct accurate predictions? (RQ2) Can our proposed model produce high-quality attribution maps?

4.1 Experimental Setup

Datasets. We conduct experiments on two benchmark datasets: (i) **ImageNet-9** [39]: It is a subset of ImageNet containing 9 classes. For each sample in the test set, ImageNet-9 also has pixel-wise masks that highlight the target object; (ii) **NICO++** [45]: NICO++ dataset contains 60 classes. The dataset has 11313, 4440, and 4440 samples for training, validation, and testing. (iii) The an8Flower [22] dataset has predefined discriminative regions of a synthetic plant model. Each class is distinguished by assigning one of six colors to either a flower or a stem, while the remaining part defaults to green. The detailed information about the datasets is in supplementary materials.

Baselines We compare COMET with representative and state-of-the-art explanation methods, including three post-hoc explanation methods, i.e., Grad-CAM [33], RT-Saliency [9], and Score-CAM [38], and four inherently explainable methods, i.e., B-cos [5], REAL-X [19], RB-AEM [13], and DiET (Distractor Erasure Tuning) [4]. We further summarize these approaches and a detailed description of the baselines is in supplementary materials.

Implementations. We use Resnet18 [14] for both the predictor and detector. To control the overall size of our model, we use the Lite R-ASPP Network model [32] with MobileNetV3 [17] as the backbone for our selector. All methods are trained from scratch for a fair evaluation and to avoid the influence of external

Table 2: Classification Performance (Accuracy±Std %)

Model	Imagenet9	NICO++	Imagenet9(noisy)	NICO++(noisy)
Resnet18 [14]	91.16 ±0.34	40.20 ±0.45	66.53 ±1.13	28.97 ±0.44
Resnet101 [14]	91.96 ±0.06	39.58 ±0.40	64.91 ±1.32	28.64 ±1.24
B-cos [5]	75.21 ±0.70	27.04 ±1.13	59.75 ±0.63	22.72 ±0.69
Realx [19]	66.13 ±2.55	21.11 ±0.11	53.97 ±0.76	8.63 ±1.24
RB-AEM [13]	78.40 ±0.34	17.93 ±0.44	50.30 ±0.98	6.84 ±1.29
DiET [4]	73.41 ±1.21	38.45 ±0.07	54.71 ±0.75	7.93 ±0.33
COMET-TD	92.82 ±1.07	**58.38** ±0.87	70.32 ±0.85	**38.30** ±0.17
COMET	**94.33** ±0.37	56.47 ±1.74	**74.81** ±0.93	33.12 ±0.10

knowledge. Images are normalized with the mean and standard deviation of the datasets. For post-hoc baselines, we apply them on the Resnet18. For inherently explainable baselines, we use Resnet18 as the backbone. We conduct a hyper-parameter search for all the methods using the validation set. For our model, the optimal coefficients are found $a = 5$, $b = 100$, and $t = 0.1$. The learning rate is set as 0.0005. The maximum epoch is set as 200 for all methods.

4.2   RQ1: Assessing Classification Accuracies

To answer RQ1, we assess the classification accuracy of our model in comparison to regular Resnet and inherently explainable baselines. Furthermore, to examine the robustness of the models, we evaluate models on the noisy version of these two datasets where we randomly remove 10% of pixels in each image. We also test COMET-TD which is a variant of our model where the detector is trained on masked-out regions along with the selector and the predictor. We conduct experiments 5 times and the average results with standard deviation. In Tab. 2, we can observe that our method outperforms other baselines on ImageNet-9 and demonstrates greater improvements on NICO++. Our model is also robust on the noisy version maintaining the highest accuracy among baselines. Other inherently explainable methods have a lower accuracy compared with regular Resnet18 and Resnet101. The reason is that they either have extra constraints that limit the prediction ability (B-cos) or would potentially discard discriminative features that can be helpful for prediction (REAL-X, RB-AEM, DiET). Our model achieves higher accuracy than Resnet18 and Resnet101. There are two possible reasons. One is that our detector module encourages all discriminative features to be seen by the predictor ensuring that our module would not perform worse than regular Resnet. Moreover, removing non-discriminative noise can improve the overall prediction performance [26].

4.3   RQ2: Assessing the Quality of Attribution Maps

To answer RQ2, we evaluate the qualities of feature attributions. We quantify the feature coverage, localization ability, fidelity, and robustness of our model.

Table 3: Area under the IOU curve (The higher the better)

Grad-CAM	Score-CAM	RT-Saliency	Bcos	DiET	COMET	COMET(ViT)	COMET-DR
15.74	17.51	25.3	39.2	11.6	**72.5**	69.5	5.83

Table 4: Localization ability test: PxAP of different methods. (The higher the better)

Post-hoc			Inherent						
Grad-CAM	Score-CAM	RT-Saliency	Realx	RB-AEM	B-cos	DiET	COMET-TD	COMET	COMET(ViT)
51.15	53.55	60.67	44.05	40.61	60.36	47.68	71.41	**76.38**	69.87

Feature Coverage: To assess whether our models can accurately cover discriminative features, we calculate the area under the IOU (intersection over union) curve for the an8Flower, which has predefined discriminative regions [22]. Additionally, we perform a sanity check [1] by training COMET-DR on randomized data, where labels are randomly assigned to images. As shown in Tab. 3, our models achieve the highest scores, indicating that they accurately cover the discriminative features. In contrast, COMET-DR, which scores the lowest, fails to generate accurate attribution maps. This significant difference from COMET demonstrates that our models produce valid attribution maps closely related to the training data.

Localization Ability: We use PxAP (Pixel Average Precision) [8] to evaluate attribution maps pixel-by-pixel against ground-truth maps and calculate the area under the precision-recall curve. The detailed definition of PxAP is in supplementary materials. As shown in Tab. 4, our proposed method outperforms other baselines. Among post-hoc methods, RT-Saliency outperforms activation-based methods (Grad-CAM and Score-CAM). The reason is that the relevance of neuron activation to the prediction is often not clear and can be heuristic. RT-Saliency directly measures the impact of altering specific features on the prediction. Realx and RB-AEM have overall low localization ability because the predictor is trained on randomly masked images that are not in the same distribution as the ones from the selector. Compared with COMET-TD, the COMET has better performance. COMET-TD would also be unstable because the detector would fit on noise in the masked-out regions, which degrades its ability to detect discriminative features. Without the influence of the selector, the pre-trained detector can effectively detect discriminative features in masked-out regions.

Fidelity: Fidelity refers to the accuracy of the attribution method in correctly identifying the features that are truly important or influential in the model's decision-making process. As a result, pixel perturbation test [4, 8, 35] is used to quantify the fidelity of our attribution method. We mask k% of pixels with the highest attribution scores, where k is varied as {5, 10, 20}. We then test the masked images with the predictor and record the accuracy. Low accuracy indicates that the masked pixels are important for prediction. Results are shown in Tab. 5. Inherently explainable methods will produce lower accuracy

Table 5: Fidelity test: Perturbed accuracy when removing pixels with top attribution scores. (Accuracy±Std %, the lower the better)

Type	Model	ImageNet-9			NICO++		
		top 5%	top 10%	top 20%	top 5%	top 10%	top 20%
Post-hoc	Grad-CAM [33]	85.20 ±0.16	79.93 ±0.42	69.96 ±0.30	29.27 ±0.91	22.20 ±0.40	14.57 ±0.28
	Score-CAM [38]	84.46 ±0.58	78.63 ±0.98	67.23 ±1.04	28.02 ±0.03	21.54 ±0.27	13.84 ±0.17
	RT-Saliency [9]	83.91 ±0.04	75.26 ±1.13	59.12 ±4.14	34.02 ±0.59	30.32 ±0.16	24.84 ±0.30
Inherent	Realx [19]	25.14 ±2.23	23.70 ±1.53	21.48 ±1.22	11.20 ±1.94	10.02 ±1.96	4.41 ±1.26
	RB-AEM [13]	67.76 ±4.12	64.53 ±4.86	56.79 ±3.73	11.92 ±0.85	11.04 ±0.31	6.24 ±0.24
	B-cos [5]	53.24 ±0.68	42.74 ±0.41	31.43 ±0.70	14.38 ±0.17	10.39 ±0.06	7.00 ±0.08
	DiET [4]	27.94 ±0.12	26.01 ±0.01	23.57 ±0.51	20.22 ±0.53	12.61 ±0.05	6.64 ±0.18
	COMET-TD	16.34 ±0.72	15.90 ±0.71	14.23 ±0.87	**6.11** ±0.18	**2.99** ±0.24	**1.87** ±0.16
	COMET	**10.64** ±1.12	**8.68** ±1.25	**2.07** ±1.24	8.97 ±0.13	4.69 ±0.27	2.74 ±0.49

Table 6: Robustness test: PxAP when randomly removing pixels or changing backgrounds (the higher the better).

Model		Noisy images	Background change
Post-hoc	Grad-CAM [33]	47.28 ±0.98	52.54 ±0.52
	Score-CAM [38]	52.90 ±0.92	53.14 ±0.85
	RT-Saliency [9]	57.53 ±1.19	54.21 ±3.06
Inherent	Realx [19]	40.93 ±0.97	42.52 ±0.53
	RB-AEM [13]	38.91 ±0.89	39.54 ±0.41
	B-cos [5]	57.28 ±0.67	57.85 ±0.59
	DiET [4]	45.20 ±1.35	42.25 ±1.23
	COMET-TD	70.30 ±1.60	73.33 ±0.56
	COMET	**73.04** ±0.85	**74.22** ±1.15

when pixels with high attribution scores are perturbed. This means that inherently explainable methods produce more faithful attribution maps, which is also shown in the previous study [19]. Our method outperforms other baselines and correctly identifies important features.

Robustness: For attributions, we also evaluate the robustness of our model. We randomly remove 20% pixels and calculate the PxAP of generated attribution maps. We test on the same test set where only the backgrounds of images are changed. Tab. 6 shows that our method has the best performance with both noisy images and background-changed images.

4.4   Further Assessment: Ablation and Hyper-Parameter Sensitivity

Aside from comparing our model with baselines, we conduct an ablation study to show the effectiveness of the detector: (i) **COMET-TD (Trainable Detector):** We train the detector with masked-out regions instead of pre-training it with full images; (ii) **COMET-w/o Detector**: We remove the detector and optimize both the selector and the predictor. In this setting, the predictor takes
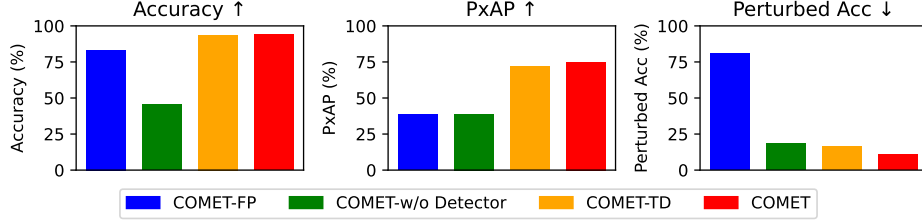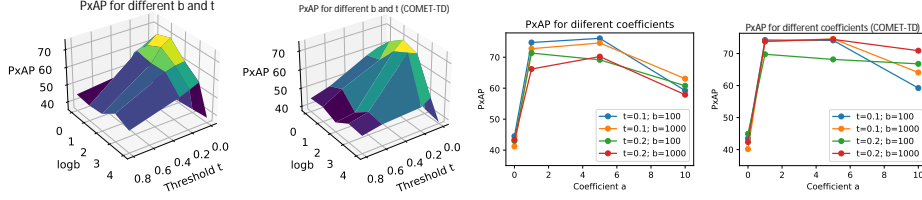
Fig. 3: Ablation study.

Fig. 4: Test for different coefficients. The left two figures are PxAP for different b and t. The right two figures are PxAP for different a.

over the task of the detector detecting any discriminative features in masked-out regions; (iii) **COMET-FP (Fixed Predictor) & w/o Detector**: In this setting, the detector is removed and the predictor is also pre-trained and fixed. We only optimize the selector.

In Fig. 3, all other models have overall lower perturbed accuracy than COMET-FP, because training the predictor and the selector jointly will have higher fidelity and find truly important pixels for the predictor. By comparing COMET-FP and COMET-w/o Detector, we can see serious degenerate cases where the accuracy drops quickly. Considering that COMET-w/o Detector makes the predictor trainable, the model should have obtained higher accuracy with more weights. However, the selector and predictor interlock each other, leading to even lower prediction accuracy. COMET improves the prediction accuracy and localization ability (PxAP) of attribution maps. COMET has better results with COMET-TD, because a pre-trained detector can effectively detect discriminative features and make the training process more stable. We perform the hyper-parameter sensitivity analysis by trying different values of a, b, and t. We test both COMET and COMET-TD. Firstly, we try different values of b and t, which control the mask regularization term. b controls stength the mask regularization term the overall training process. t is a threshold value and larger t indicates more pixels can be accessed by the predictor. As shown in Fig. 4, the PxAP of both COMET and COMET-TD reaches the top values when $b \in \{100, 1000\}$ and $t \in \{0.1, 0.2\}$. We further test different values of a. COMET reaches the maximum value when $a = 5$, $b = 100$ and $t = 0.1$. The PxAP decreases dramatically when $a = 0$ (No control on masked-out regions). This shows that the control on masked-out regions can improve the attribution maps.
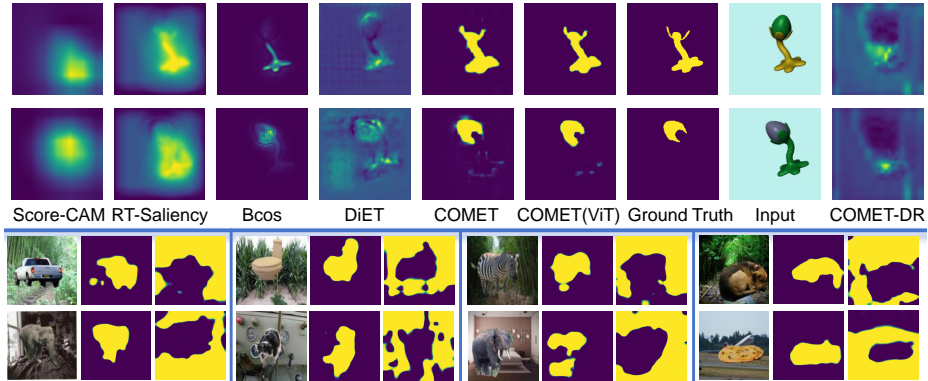
Fig. 5: Top (an8Flower): Attribution maps comparison; Bottom (BAM): Attribution maps when model trained on the object (left) or scene (right) labels.

4.5   Visualization

As shown in Fig. 5, COMET highlights the color-assigned, discriminative part of the plant while excluding the non-discriminative (green) part. This shows that our model can accurately cover discriminative features instead of highlighting the whole object. We also train our model on the BAM dataset, synthesizing objects and scenes [40]. Using the same training set, we train two models: one with object labels and another with scene labels. As shown in Fig. 5 (bottom), the object-labeled model focuses on the object, while the scene-labeled model emphasizes the scenes.

5   Discussion

Despite the effectiveness, we recognize two limitations. The sparsity of the mask can be challenging to uniformly apply due to varying target sizes across images. Secondly, the model's complexity leads to longer training time. However, this is a one-time investment, and our model has the shortest inference time. We compare computation time and report model size in the supplementary materials.

6   Conclusion

Our proposed model COMET, effectively addresses the incompleteness problem by introducing an additional objective that penalizes the presence of discriminative features in the masked-out regions. This enhances the comprehensiveness of feature selection, resulting in accurate and complete attribution maps. An independent detector is introduced to detect discriminative features in masked-out regions. The independent detector can effectively break the interlocking situation of the selector and the predictor. Experiments demonstrate that our method can generate high-quality attribution maps while making accurate predictions.

Acknowledgements

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. Advances in Neural Information Processing Systems **33**, 700–712 (2020)
3. Akhtar, N.: A survey of explainable ai in deep visual modeling: Methods and metrics. arXiv preprint arXiv:2301.13445 (2023)
4. Bhalla, U., Srinivas, S., Lakkaraju, H.: Discriminative feature attributions: Bridging post hoc explainability and inherent interpretability. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
5. Böhle, M., Fritz, M., Schiele, B.: B-cos networks: Alignment is all we need for interpretability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10329–10338 (2022)
6. Carter, B., Jain, S., Mueller, J.W., Gifford, D.: Overinterpretation reveals image classification model pathologies. Advances in Neural Information Processing Systems **34**, 15395–15407 (2021)
7. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: An information-theoretic perspective on model interpretation. In: International conference on machine learning. pp. 883–892. PMLR (2018)
8. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3133–3142 (2020)
9. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. Advances in neural information processing systems **30** (2017)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Shieber, S., Waldo, J., Weinberger, D., et al.: Accountability of ai under the law: The role of explanation. DEVELOPMENT IN PRACTICE **663**, 663–65 (2007)
12. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
13. Ganjdanesh, A., Gao, S., Huang, H.: Interpretations steered network pruning via amortized inferred saliency maps. In: European Conference on Computer Vision. pp. 278–296. Springer (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

15. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
16. House, W.: Blueprint for an ai bill of rights: Making automated systems work for the american people. NIMBLE BOOKS (2022)
17. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
18. Jalwana, M.A., Akhtar, N., Bennamoun, M., Mian, A.: Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16327–16336 (2021)
19. Jethani, N., Sudarshan, M., Aphinyanaphongs, Y., Ranganath, R.: Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In: International Conference on Artificial Intelligence and Statistics. pp. 1459–1467. PMLR (2021)
20. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing **30**, 5875–5888 (2021)
21. Lundstrom, D.D., Huang, T., Razaviyayn, M.: A rigorous study of integrated gradients method and extensions to internal neuron attributions. In: International Conference on Machine Learning. pp. 14485–14508. PMLR (2022)
22. Oramas, J., Wang, K., Tuytelaars, T.: Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In: International Conference on Learning Representations (2019)
23. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
24. Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V.I., Mehra, A., Ordonez, V., Saenko, K.: Black-box explanation of object detectors via saliency maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11443–11452 (2021)
25. Ramaswamy, H.G., et al.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 983–991 (2020)
26. Rao, S., Böhle, M., Parchami-Araghi, A., Schiele, B.: Using explanations to guide models. arXiv preprint arXiv:2303.11932 (2023)
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
28. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
29. Regulation, G.D.P.: General data protection regulation (gdpr). Intersoft Consulting, Accessed in October **24**(1) (2018)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

32. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
34. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
35. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. Advances in neural information processing systems **32** (2019)
36. Srinivas, S., Fleuret, F.: Rethinking the role of gradient-based attribution methods for model interpretability. In: International Conference on Learning Representations (2020)
37. Stalder, S., Perraudin, N., Achanta, R., Perez-Cruz, F., Volpi, M.: What you see is what you classify: Black box attributions. Advances in Neural Information Processing Systems **35**, 84–94 (2022)
38. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
39. Xiao, K.Y., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. In: International Conference on Learning Representations (2020)
40. Yang, M., Kim, B.: Benchmarking Attribution Methods with Relative Feature Importance. CoRR **abs/1907.09701** (2019)
41. Yang, P., Akhtar, N., Wen, Z., Mian, A.: Local path integration for attribution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3173–3180 (2023)
42. Yoon, J., Jordon, J., van der Schaar, M.: Invase: Instance-wise variable selection using neural networks. In: International Conference on Learning Representations (2018)
43. Yu, M., Chang, S., Zhang, Y., Jaakkola, T.S.: Rethinking cooperative rationalization: Introspective extraction and complement control. In: Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing. Association for Computational Linguistics (2019)
44. Yu, M., Zhang, Y., Chang, S., Jaakkola, T.: Understanding interlocking dynamics of cooperative rationalization. Advances in Neural Information Processing Systems **34**, 12822–12835 (2021)
45. Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., Cui, P.: Nico++: Towards better benchmarking for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16036–16047 (2023)
46. Zhang, Y., Barzilay, R., Jaakkola, T.: Aspect-augmented adversarial networks for domain adaptation. Transactions of the Association for Computational Linguistics **5**, 515–528 (2017)
47. Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S., Bianchi, M.T.: Learning sleep stages from radio signals: A conditional adversarial architecture. In: International conference on machine learning. pp. 4100–4109. PMLR (2017)
48. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)