IMPROVING FEATURE ALIGNMENT IN CONVNETS USING CONTRASTIVECAMS AND CORE-FOCUSED CROSS-ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the ubiquity of modern deep learning, accurate explanations of network predictions remain largely elusive. HiResCAM is a popular interpretability technique used to visualize attention maps (i.e., regions-of-interest) over input images. In this paper, we theoretically show a limitation of HiResCAM: the HiResCAMs for a given input are not uniquely determined, allowing an arbitrary spurious shift by a common matrix M while corresponding to the same prediction. We further propose *ContrastiveCAMs*, which are invariant to the spurious shift M hence improving robustness of explanations, while additionally providing granular class-versus-class explanations. With the additional granular explanations, experiments reveal that networks often focus on regions unrelated to the class label. To address this issue, we leverage the knowledge of core image regions and propose *Core-Focused Cross-Entropy*, an extension of cross entropy, which encourages attention on core regions while suppressing unrelated regions, improving feature alignment. Experiments on Hard-ImageNet and Oxford-IIIT Pets show that ContrastiveCAM provides more faithful attention maps and our method effectively improves feature alignment by primarily extracting predictive performance from core image regions.

1 INTRODUCTION

The vast applications of convolutional neural networks in safety-critical domains such as medical imaging (Kc et al., 2021; Rajpurkar et al., 2017), forensic investigation (Murthy and Siddesh, 2023) and self-driving (Kim and Canny, 2017) make accurate (a.k.a faithful) interpretations of their predictions paramount (Haufe et al., 2024). Approaches to explain predictions include feature-attribution based interpretability techniques (Zhou et al., 2016; Selvaraju et al., 2017; Draelos and Carin, 2020), input-based interpretability with saliency maps (Simonyan et al., 2013; Smilkov et al., 2017), and more recently, mechanistic interpretability for image circuit discovery (Olah et al., 2020).

In addition to faithful interpretability, ensuring that only target-relevant (a.k.a. core) regions influence model predictions is a critical determination to make. A model-agnostic approach for evaluating the impact of core regions involves input ablation experiments as introduced in recent work on Core Risk Minimization (Singla et al., 2022; Moayeri et al., 2022). Images are modified to systematically corrupt core regions, following which the change in performance is reported. Singla and Feizi (2022) demonstrate that both convolutional and transformer-based architectures are vulnerable to learning non-core regions of the input, caused by features like co-occurring backgrounds. These encourage learning 'tricks' – shortcuts to learning that improve in-distribution accuracy while inhibiting generalization over core features (Geirhos et al., 2020). A concrete example of shortcut learning is illustrated within the introduction of Invariant Risk Minimization (Arjovsky et al., 2020).

In this work, we develop and leverage faithful interpretability to encourage feature alignment in convolutional models. We theoretically observe that HiResCAMs (Draelos and Carin, 2020) may not explain true factors that contribute towards predictions as a consequence of *softmax* activation. Specifically, we prove that HiResCAMs are not uniquely determined and admit arbitrary, spurious shifts by a common matrix M while corresponding to the same prediction (Theorem 3.2). This spurious shift from M can, in principle, completely corrupt HiResCAM explanations. To remove this redundancy, we propose *ContrastiveCAMs* (Definitions 3.3, 3.4), resulting in attention maps that

are invariant to the aforementioned spurious shift while additionally providing granular class-versus-class explanations. Using class-versus-class comparisons, we experimentally reveal circumstances wherein different comparisons leverage different regions to base their predictions. Further, these differing regions do not always correspond to core regions of the input image, i.e., there are spurious contributions. We demonstrate that cross entropy loss encourages leveraging these unrelated regions, especially in settings where the target represents a small portion of the image (Section 4.1). Finally, we propose a modification to cross-entropy, termed *Core-Focused Cross-Entropy* (Definition 4.5), which: a) suppresses user-specified non-core regions despite the presence of spurious factors, and b) generates contrast within user-specified target regions to solve for the underlying classification task. This improves feature alignment by encouraging the model to learn target-relevant features only.

We demonstrate the effectiveness of our proposed method by reporting experimental results in multiclass, multiple-class, and binary classification settings. We supplement this evidence by showing that core-focused models may be trained competitively even with coarse or auto-generated masks, and that they outperform backbones trained using cross-entropy in downstream segmentation tasks.

1.1 Related Work

Feature Attribution in Convolutional Networks. A prominent family of interpretability techniques stems from the seminal CAMs (short for Class Activation Mappings) (Zhou et al., 2016) literature. CAMs help identify regions-of-interest in the form of attention maps. It's success led to the introduction of a vast set of derivative works, that extend CAMs in various ways (Selvaraju et al., 2017; Chattopadhay et al., 2018; Wang et al., 2020; Draelos and Carin, 2020).

Representation Learning. Arjovsky et al. (2020) introduces the notion of predictors that learn feature representations that are invariant to spurious factors. Bau et al. (2017) quantifies the interpretability of learned representations in convolutional models by evaluating hidden units within convolutional layers on segmentation tasks. Recently, Zou et al. (2023) motivates neuroscience-inspired top-down approached for inducing interpretability. It encourages the analysis of representations (representation reading) and it's subsequent modification (representation control).

Feature Alignment. Spurious factors in images encourage extracting predictions from unrelated regions, termed *shortcuts*, and are discussed extensively by Geirhos et al. (2020). Feature alignment seeks to ensure predictions are made using relevant features only, and is deeply connected with robustness in neural networks (Wang, 2023). Preventing shortcut learning is thus a crucial goal of feature alignment. Approaches to alignment include region masking (Kc et al., 2021), tiered training (Aniraj et al., 2023), and regularization via saliency maps (Ismail et al., 2021), each having an empirical focus. For a thorough exposition to recent advancements and challenges in interpretability-guided feature alignment, we direct the reader to Weber et al. (2023) and Gao et al. (2024).

2 Preliminaries

Notation. We denote vectors using bold lowercase letters (e.g., $\mathbf{v}$), matrices using uppercase letters (e.g., M), and tensors using bold uppercase letters (e.g., $\mathbf{T}$), with partial indexing implying selection of the subtensor across the remaining subsequent dimensions (e.g., $\mathbf{T}_i \in \mathbb{R}^{b\times c}$ for $\mathbf{T} \in \mathbb{R}^{a\times b\times c}$). We use the operator $\odot$ to represent elementwise multiplication, and define $[C] := \{1, 2, \ldots, C\}$.

Setup. In this paper, we consider image classification tasks. The dataset $\mathcal{D} = \{(\mathbf{X}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{n}$ contains image-label pairs where images are represented using rank-3 tensors $\mathbf{X}$ consisting of two spatial dimensions and one channel dimension, and labels are one-hot vectors $\mathbf{y} \in \mathbb{R}^C$, where C denotes the total number of classes in the dataset. A neural network f is trained to learn the relation between the images $\mathbf{X}$ and labels $\mathbf{y}$. The output of f contains C *logits*: f_c, $c \in [C]$. Let $\sigma(\cdot)$ be the softmax function, $\tilde{f}(\mathbf{X}) = \sigma(f(\mathbf{X}))$ is then interpreted as the class-specific *probability predictions*. The standard training procedure is to optimize a cross-entropy loss function so that $\tilde{f}(\mathbf{X})$ matches the label $\mathbf{y}$ as closely as possible for each training image-label pair.

In prominent approaches such as VGG (Simonyan and Zisserman, 2015), ResNet (He et al., 2016) & ViT (Dosovitskiy et al., 2020), the neural network f mainly consists of two consecutive parts, a

backbone module g followed by a classifier h: $f = h \circ g$. In this paper, we focus on convolutional neural networks, i.e., the backbone g is convolutional. We denote the output of the backbone g as $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1 \times d_2}$, termed as feature embedding (a.k.a. feature maps) of the image, where d_0 is the number of features (a.k.a. channels) and d_1 and d_2 are the spatial dimensions of the final convolutional layer. The feature embedding $\mathbf{A}$ is then reduced to a vector $\mathbf{z}$, either by flattening $\mathbf{z} = vec(\mathbf{A})$, or by Global Average Pooling (GAP). $\mathbf{z}$ is then processed by the classifier h, which outputs the logits f, that are passed through *softmax* to obtain the class prediction vector, denoted $\tilde{f}$.

The recent trend is that the classifier h becomes as simple as a single layer, such as in ConvNext (Liu et al., 2022), ViT (Dosovitskiy et al., 2020), EfficientNet (Tan and Le, 2019), ResNet (He et al., 2016) & DenseNet (Iandola et al., 2014):

$$h(\mathbf{z}) = W\mathbf{z} + \mathbf{b}, \tag{1}$$

This simplification of h is largely due to the fact that the backbone g, which encapsulates the bulk of the model's predictive power, extracts high quality and comprehensive features $\mathbf{A}$, based on which a single layer is enough to obtain accurate final predictions. In this paper, we assume that the classifier is of the form in Eq. (1).

HiResCAMs. HiResCAMs (short for High-Resolution Class Activation Maps), introduced in (Draelos and Carin, 2020), is a method designed to provide interpretable explanations of convolutional neural networks. It renders the contribution of each spatial location in an image to the final logit output f_c, thereby revealing which regions are most critical to the models prediction. Specifically, given a feature embedding $\mathbf{A}$ of an image $\mathbf{X}$ and a class index $c \in [C]$, the HiResCAM is defined as:

$$\mathbf{CAM}_c^{\text{HiRes}} = \sum_{j=1}^{d_0} (\nabla_{\mathbf{A}_j} f_c) \odot \mathbf{A}_j, \qquad \mathbf{CAM}_c^{\text{HiRes}} \in \mathbb{R}^{d_1 \times d_2} \tag{2}$$

$\mathbf{CAM}_c^{\text{HiRes}}$ shares spatial dimensions with the backbone output $\mathbf{A}$. Each element within $\mathbf{CAM}_c^{\text{HiRes}}$ represents a contribution to the logit output f_c from a corresponding patch within the original image. A higher absolute value implies a greater contribution.

HiResCAMs have been widely used for incorporating explainability in a variety of tasks, such as CT scan abnormality classification (Draelos and Carin, 2022), malware visualization (Brosolo et al., 2025), coffee leaf rust classification (Chavarro et al., 2024), counterfeit banknote detection (Pachón et al., 2023) & flow estimation (Chen and Wu, 2025).

Particularly, for single-layer classifiers h, Draelos and Carin (2020) show that the expression of HiResCAMs, Eq. (2), can be simplified and has the following close connection with output logits f_c:

$$f_c(\mathbf{X}) = \sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}_{c,i,j}^{\text{HiRes}}(\mathbf{X}) + \mathbf{b}_c, \qquad c \in [C]. \tag{3}$$

Each logit f_c is the summation of the HiResCAM over its spatial dimensions, up to a scalar $\mathbf{b}_c$.

3 Contrastive Class Activation Maps

In this section, we first discuss the theoretical limitations of HiResCAM in explaining model predictions, and then introduce a surrogate method, ContrastiveCAM, which offers more faithful and class-specific explanations.

HiResCAMs Admit Spurious Shifts. A key observation is that HiResCAMs are only related to *logits* f, not *probability predictions* $\tilde{f} = \sigma(f)$ belonging to each class, see Eq. (3). The drawback is that, for the same probability prediction $\tilde{f}$, there are infinitely many possible logit outputs f, hence infinitely many HiResCAMs, each of which explain the same prediction differently. This drawback arises intrinsically from the nature of the *softmax* function.

Proposition 3.1 (Contrastiveness of *softmax*)**.** *The softmax function is invariant to a universal shift of all its input components:*

$$\sigma(\mathbf{x}) = \sigma(\mathbf{x} + a\mathbf{1}_C) \qquad \forall \mathbf{x} \in \mathbb{R}^C,\ a \in \mathbb{R} \tag{4}$$

Proof. All proofs are deferred to Appendix A. □

This invariance to $a \in \mathbb{R}$ is amplified to a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ when assessing HiResCAMs.

Theorem 3.2. *HiResCAM explanations* $\mathbf{CAM}^{\text{HiRes}} \in \mathbb{R}^{C \times d_1 \times d_2}$ *corresponding to probability predictions* $\tilde{f}(\mathbf{X}) \in \mathbb{R}^C$ *are not uniquely determined, admitting a universal shift of class-level explanations* $\mathbf{CAM}_c^{\text{HiRes}}$ *by an arbitrary matrix* $M \in \mathbb{R}^{d_1 \times d_2}$ $\forall c \in [C]$.

$$\tilde{f}(\mathbf{X}) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}_{:,i,j}^{\text{HiRes}} + \mathbf{b}\right) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \overline{\mathbf{CAM}}_{:,i,j}^{\text{HiRes}} + \mathbf{b}\right) \quad \forall M \in \mathbb{R}^{d_1 \times d_2} \tag{5}$$

Where $\overline{\mathbf{CAM}}^{\text{HiRes}}$ *is defined as:*

$$\overline{\mathbf{CAM}}_c^{\text{HiRes}} := \mathbf{CAM}_c^{\text{HiRes}} + M \qquad \forall c \in [C] \tag{6}$$

Thus explanations from HiResCAMs are accurate only upto a summand M which is unknown. These explanations may be misleading, and *fail to guarantee a faithful interpretation* of the model prediction. An example of such a misinterpretation is illustrated in Figure 1.

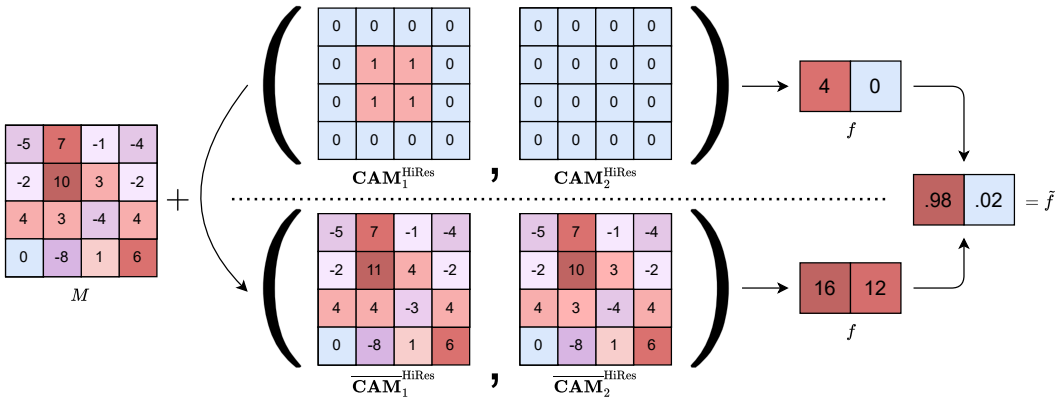


Figure 1: Shifting $\mathbf{CAM}^{\text{HiRes}}$ by arbitrary matrix M results in a change to explanations $\overline{\mathbf{CAM}}^{\text{HiRes}}$ which subsequently changes the corresponding logit vector. However, the model's final prediction probabilities are identical and remain unchanged.

To remove this redundancy, we define a contrastive representation of HiResCAMs, which recovers faithful attention maps at the class probability level.

Definition 3.3 (ContrastiveCAMs)**.** Given a set of classes $[C]$ with c_t being the index of the target class for a given image, ContrastiveCAM is defined as follows:

$$\mathbf{CAM}_{c_t}^{\text{Cntrst}} := \left\{\mathbf{CAM}_{(c_t,c')}^{\text{Cntrst}} : c' \in [C] \setminus c\right\}, \quad \mathbf{CAM}_{(c_t,c')}^{\text{Cntrst}} := \mathbf{CAM}_{c_t}^{\text{HiRes}} - \mathbf{CAM}_{c'}^{\text{HiRes}} \tag{7}$$

Further, we also reconstruct single-class interpretations of ContrastiveCAMs:

Definition 3.4 (Class-Reconstructed ContrastiveCAMs)**.** Given a set of classes $[C]$ with c_t being the index of the target class for a given image, reconstructed ContrastiveCAMs are defined as follows:

$$\mathbf{CAM}_{c_t}^{\text{Recon}} := \frac{1}{C}\sum_{c=1}^{C} \mathbf{CAM}_{(c_t,c)}^{\text{Cntrst}} = \mathbf{CAM}_{c_t}^{\text{HiRes}} - \frac{1}{C}\sum_{c=1}^{C} \mathbf{CAM}_c^{\text{HiRes}} \tag{8}$$

$\mathbf{CAM}_{c_t}^{\text{Recon}}$ thus removes redundancy $R = -1/C \cdot \sum_{c=1}^{C} \mathbf{CAM}_c^{\text{HiRes}}$. We report the ratio of redundancy to the original explanation as $\gamma = \|R\|_F / \|\mathbf{CAM}_{c_t}^{\text{HiRes}}\|_F$ for various datasets in Table 1.

Crucially, ContrastiveCAMs are invariant to spurious contributions as exposed by Theorem 3.2.

Theorem 3.5 (ContrastiveCAMs are M-invariant)**.** *Let* $\mathbf{CAM}^{\text{HiRes}}$ *and* $\overline{\mathbf{CAM}}^{\text{HiRes}}$ *be two HiResCAMs corresponding to probability predictions* $\tilde{f}(\mathbf{X}) \in \mathbb{R}^C$ *such that:*

$$\overline{\mathbf{CAM}}_c^{\text{HiRes}} = \mathbf{CAM}_c^{\text{HiRes}} + M \qquad \forall c \in [C] \tag{9}$$

Then, for every $M \in \mathbb{R}^{d_1 \times d_2}$*, it holds that:*

$$\mathbf{CAM}^{\text{Cntrst}} = \overline{\mathbf{CAM}}^{\text{Cntrst}} \quad \text{and} \quad \mathbf{CAM}^{\text{Recon}} = \overline{\mathbf{CAM}}^{\text{Recon}} \tag{10}$$

Class-versus-Class Explanations. While explanations from the CAM-family only involve visualizing f_{c_t}, softmax activation uses every logit in computing class probabilities. Making inferences based on individual logits may thus misinterpret the internal model state, as the training objective induced by cross-entropy loss over softmax activation is to maximize the **difference between class logits**, see Eq. (44). We demonstrate the value of additional granularity provided by pairwise explanations by reporting observations on a three-class subset of Hard-ImageNet in Figure 2.

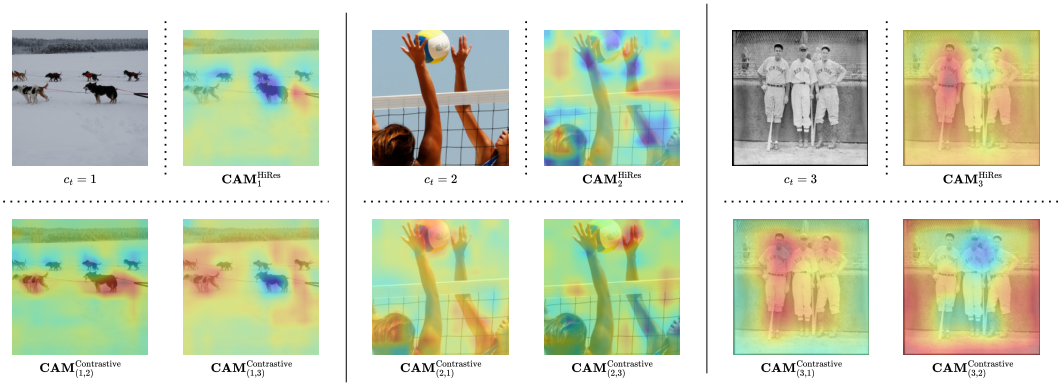


Figure 2: We plot ContrastiveCAM and HiResCAM explanations on a ResNet-18 model trained to classify: ('dog sled', 'volleyball', 'baseball player'), ordered by label index. ContrastiveCAMs reveal circumstances wherein: a) regions that contribute towards prediction are hidden by HiResCAMs, and b) differing parts of the image contribute towards various class-versus-class predictions.

From Figure 2, we also observe that the model often leverages irrelevant regions (e.g., environmental cues), to inform predictions. Following Moayeri et al. (2022), we refer to these regions as *non-core regions*. In principle, *core regions* are those that causally influence the prediction (modification of this region could mean the ground truth itself may change), while *non-core regions* represent spurious correlations – modifications to these regions do not change the ground truth labels.

Table 1: ContrastiveCAM explanations bifurcated by core-region maps across various datasets. The average contributions of core / non-core regions and ratio of redundancy removed is reported below.

Dataset	Core (↑)	Non-Core (↓)	$^{\text{Core}}/_{\text{Total}}$ (↑)	Redundancy (γ)	Accuracy (%)
Hard-ImageNet	14.817	**42.138**	.2601	.201	95.73
Oxford-IIIT Pets	**3.925**	2.150	.6461	.367	99.34
PASCAL VOC	1.581	**1.719**	.4791	$-$[1]	87.32

This undesired influence is consistently observed, as evidenced by high overall non-core contribution in Table 1 above. Despite strong accuracy, large contributions arise from non-core regions.

4 Learning with ContrastiveCAMs

The dependency on non-core regions observed above is evidence of misalignment, which inhibits generalization. In this section, we prove a desirable theoretical property of ContrastiveCAMs and leverage it to incorporate interpretability within model optimization, mitigating this weakness.

Specifically, we prove that any *input-dependent* change to probability predictions $\tilde{f}$ (e.g., caused by updating model weights) is precisely reflected by a proportionate change to $\mathbf{CAM}^{\text{Cntrst}}_{c_t}$.

Proposition 4.1 (Correctness of ContrastiveCAMs)**.** *Softmax-activated class probabilities $\tilde{f}$ can be expressed as a direct function of ContrastiveCAMs and the bias vector.*

$$\tilde{f}_{c_t}(\mathbf{X}) = \left(\sum_{c=1}^{C} \exp\left(\mathbf{b}_c - \mathbf{b}_{c_t} - \sum \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} \right) \right)^{-1} \qquad \forall c_t \in [C] \tag{11}$$

Where $\mathbf{CAM}^{\text{Cntrst}}_{(c_t,c_t)} = \mathbf{0}_{d_1 \times d_2}$.

By zero-ing the final bias vector (i.e., $\mathbf{b} := \mathbf{0}_C$ for h *only*), we can precisely disassociate the role of specific regions in computing cross-entropy. We leverage this property to study feature misalignment, and later in our proposed modification of cross-entropy to penalize the use of non-core regions.

4.1 Cross-Entropy Can Motivate Feature Misalignment

To encode core-region information, for each sample from our dataset of size N, we extend dataset $\mathcal{D}$ by specifying a binary mask H, which indicates whether or not downsampled regions from the input image may be used to determine the prediction.

$$\mathcal{D} := \{(\mathbf{X}^{(i)}, (H^{(i)}, \mathbf{y}^{(i)}))\}_{i=1}^{N} \quad \text{where} \quad H_{jk} := \begin{cases} 1 & \text{region contains target} \\ 0 & \text{region doesn't contain target} \end{cases} \forall j,k \in [d_1],[d_2]$$

We can restate cross-entropy as a function of ContrastiveCAMs and core-region information in $\mathcal{D}$.

Proposition 4.2. *Given bias-free classifier h, we can precisely associate the impact of specific regions, encoded by binary mask H, to the computation of cross-entropy loss.*

$$\mathcal{L}_{\text{CE}}(f(\mathbf{X}), \mathbf{y}, H) = \log\left(\sum_{c=1}^{C} \exp\left(-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} - \sum (1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} \right) \right) \tag{12}$$

Remark 4.3. *Equivalently, we disassociate the logit and use the standard cross-entropy formulation:*

$$\mathcal{L}_{\text{CE}}(f(\mathbf{X}), \mathbf{y}, H) = \mathcal{L}_{\text{CE}}\left(\sigma\left(-\sum_{i=1,j=1}^{d_1,d_2} \underbrace{H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,:),i,j}}_{\text{core}} + \underbrace{(1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,:),i,j}}_{\text{non-core}} \right), \mathbf{y} \right) \tag{13}$$

We observe from Proposition 4.2 that cross-entropy loss does not inherently favor using the core or non-core regions for classification. Provided the prediction is accurate with high confidence, error remains low. This presents a theoretical basis for feature misalignment in convolutional networks.

Scale-Sensitivity of Convolutional Approaches. In training classification models, an implicit assumption is that the strongest indicator of the class label is the target itself (i.e., the core regions). From Table 1, we observe through the significant influence of non-core regions that this assumption does not universally hold. In cases where the target is far from the camera, as commonly observed in Hard-ImageNet, the emphasis is placed on **learning the best non-core surrogate to the actual target**, rather than obtaining an accurate feature representation using just the fewer relevant regions.

Learning a non-core surrogate does reduce cross-entropy loss, but at the cost of misrepresenting the underlying classification target, thus inducing feature misalignment. The model should, through the course of training, distinguish and ignore non-core regions in determining the final prediction.

This leads us to propose an alignment-motivated constraint to empirical risk minimization.

Definition 4.4 (Core-Constrained Risk Minimization)**.**

$$\mathcal{R}_{\text{CCRM}}(f) := \mathbb{E}_{(\mathbf{X},(H,\mathbf{y}))\sim\mathcal{D}}\left[\ell(\tilde{f}(\mathbf{X}), \mathbf{y}) \right] \quad \text{s.t.} \quad \sum_{c=1}^{C} \left\| (1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} \right\| = 0 \tag{14}$$

Where $\ell(f(\mathbf{X}), \mathbf{y}) = \mathbb{1}(\arg\max(\tilde{f}(\mathbf{X})) \neq \arg\max(\mathbf{y}))$ is 0/1 loss for the multiclass setting.

4.2 CORE-FOCUSED CROSS-ENTROPY

We have shown that cross-entropy motivates generating predictions using either core or non-core features. To correct this, we propose Core-Focused Cross-Entropy, which penalizes the contribution from non-core regions to the final classification.

Definition 4.5 (Core-Focused Cross-Entropy)**.** We integrate masked region suppression to the definition of cross-entropy using the following formulation:

$$\mathcal{L}_{\text{CFCE}}(f(\mathbf{X}), \mathbf{y}, H) := \log\left(\sum_{c=1}^{C} \exp\left(-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} + \sum (1-H) \odot |\mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}|\right)\right) \tag{15}$$

We can show that the above loss function is consistent with our constrained optimization objective.

Theorem 4.6 (Consistency of Core-Focused Cross-Entropy)**.** *A sequence of predictors f_n that converges to the optimal $\mathcal{R}_{\text{CFCE}}$-risk also converges to the Bayes-optimal $\mathcal{R}_{\text{CCRM}}$-risk. Equivalently, in the realizable setting, $\mathcal{L}_{\text{CFCE}}$ is classification-calibrated.*

$$\mathcal{R}_{\text{CFCE}}(f_n) \to \mathcal{R}^*_{\text{CFCE}} \implies \mathcal{R}_{\text{CCRM}}(f_n) \to \mathcal{R}^*_{\text{CCRM}} \tag{16}$$

Where $\mathcal{R}_{\text{CFCE}}(f)$ is defined as:

$$\mathcal{R}_{\text{CFCE}}(f) := \mathbb{E}_{(\mathbf{X},(H,\mathbf{y}))\sim\mathcal{D}}\left[\mathcal{L}_{\text{CFCE}}(f(\mathbf{X}), \mathbf{y}, H)\right] \tag{17}$$

Divergence Regularization. Using ContrastiveCAMs, we observe a tendency for cross-entropy to only generate contrast in regions where feature differences are prominent within the training set. Successful test predictions rely on the prominence of the same set of differing features even if there exist subtleties in the training set that can be used to offer more nuanced classifications. We thus propose regularization by minimizing divergence between target mask H and $\mathbf{CAM}^{\text{Cntrst}}_c$. This encourages contrast for every region in which the target is present, even when the difference is subtle.

Definition 4.7 (Regularized Core-Focused Cross-Entropy)**.** We regularize $\mathcal{L}_{\text{CFCE}}$ to encourage contrast over the entire target region using KL Divergence:

$$\mathcal{L}_{\text{RCFCE}}(f(\mathbf{X}), \mathbf{y}, H) := \mathcal{L}_{\text{CFCE}} + \frac{\lambda_1}{C-1} \sum_{c\in[C]\setminus c_t} D_{KL}\left(\sigma(\lambda_2 H) \,||\, \sigma\left(\lambda_3 \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)\right) \tag{18}$$

The divergence term motivates similarity in the *shape* of ContrastiveCAMs to H. The normalizing behavior of softmax, analogous to its effect on the logits, means that absolute scale is invariant; that information comes exclusively from $\mathcal{L}_{\text{CFCE}}$.

Supplemental formulations and adaptations of core-focused optimization are deferred to Appendix B.

5 EXPERIMENTS

For our experiments, we evaluate the performance of ResNet-50 with a set of interpretability-motivated modifications. These are detailed in Appendix C. For consistency, we include baselines with (denoted by 'w/ Arch') and without these modifications. We initialize each training run on ImageNet pre-trained weights, and report fine-tuning performance.

Datasets. We present training results for Oxford IIIT-Pets (Parkhi et al., 2012), Hard-ImageNet (Moayeri et al., 2022), and the Semantic Boundaries Dataset (Hariharan et al., 2011). These datasets span image classification tasks with binary, multiclass & multilabel targets. In addition to reporting raw prediction performance, we also report intersection-over-union (IoU) scores, indicating the overlap between ground-truth core regions and those used by the models for classification.

5.1 HARD-IMAGENET

Hard-ImageNet (Moayeri et al., 2022) is a subset of ImageNet (Deng et al., 2009) that only contains classes that have been observed to use spurious features to inform predictions (Singla and Feizi, 2022).

The core regions from these classes typically constitute a minority of the overall image (13.96% on average), lending further evidence to the scale-sensitivity of convolutional models (Section 4.1).

To evaluate the performance of models using core regions only, Moayeri et al. (2022) introduces an evaluation suite that reports a) accuracy when core regions are removed from the image using segmentation masking, bounding-box masking and tiling over the foreground; b) *relative foreground sensitivity* (RFS) which evaluates performance degradation under corruption of the foreground; and c) saliency alignment measured by intersection over union of core masks to regions used for prediction.

Table 2: Hard-ImageNet benchmarks on finetuned ResNet-50 models trained using varying approaches. Models trained using our proposed core-focused loss functions show significant improvement across all evaluations, at the cost of some un-ablated performance.

Method	**Accuracy under Core-Region Ablation (%)**				**RFS** (↑)	**GradCAM IoU** (↑)	**Contrastive-CAM IoU** (↑)
	None (↑)	**Gray Mask** (↓)	**Gray BBOX** (↓)	**Tile** (↓)			
Cross-Entropy	94.25	75.94	69.39	67.38	-0.18	18.44	–
CORM (Singla et al., 2022)	92.91	76.20	69.12	68.32	-0.08	20.43	–
DFR (Kirichenko et al., 2022)	**94.39**	73.53	67.51	66.71	-0.27	18.39	–
CORM + DFR	91.31	72.59	63.64	63.90	-0.23	20.35	–
CE w/ Arch	93.69±0.77	76.53±2.15	72.49±2.19	71.02±2.4	-0.23±0.05	16.25±14.07	30.27±3.99
CFCE (Ours)	90.53±0.69	**41.78**±1.49	**31.66**±1.26	**34.31**±1.04	.224±0.10	18.88±1.13	89.22±0.31
CFCE + KL (Ours)	90.35±1.59	45.49±5.15	37.07±4.57	39.47±4.12	**.236**±0.10	**51.52**±1.07	**93.39**±0.11

IoU for this benchmark was computed using GradCAMs (Selvaraju et al., 2017) only for consistency with baselines, as GradCAMs have been shown to present unfaithful explanations (Draelos and Carin, 2020). We thus include additional evaluations using ContrastiveCAMs for core-focused models. We also qualitatively evaluate improvements using core-focused approaches in Figure 3 below.

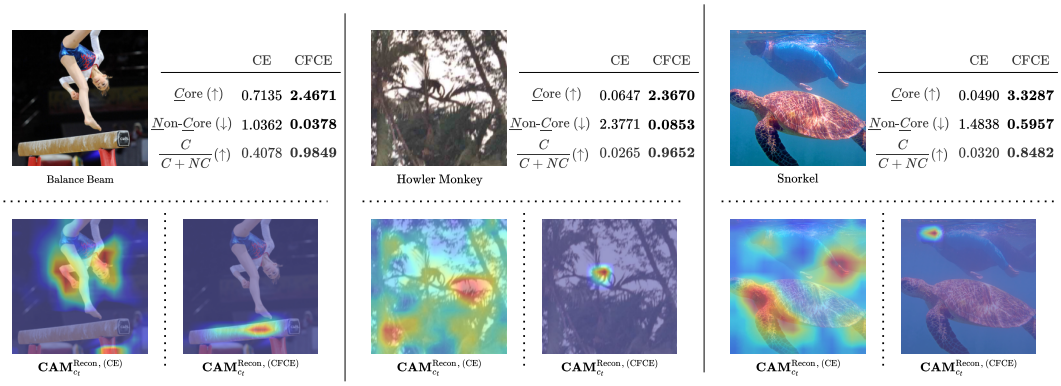


Figure 3: Models trained using CFCE exhibit suppressed contributions from non-core regions.

5.2 Oxford IIIT-Pets

The Oxford IIIT-Pets dataset contains images of 37 breeds of cats and dogs, paired with segmentation trimaps that denote the foreground and background regions within the image. In the binary setting, the objective is to classify cats and dogs; individual breed labels are merged. This creates a class imbalance (4978 dogs to 2371 cats), however no training modifications are made to account for this. There is virtually no class imbalance in the multiclass setting.

Applicability of Approximate Masks. Core-region masks H have a smaller resolution compared to input $\mathbf{X}$ as a consequence of the convolutional backbone g. Thus, in the absence of ground-truth core-region masks, approximate pixel-level masks or weaker supervision such as bounding boxes can be used to effectively suppress contributions from non-core regions. We demonstrate this empirically through competitive alignment achieved both with auto-generated masks obtained using Segment Anything (Kirillov et al., 2023) (SAM), and with weaker supervision via bounding boxes (BBOX).

Method	Core Region Masks	Binary				Multiclass			
		Accuracy (%)		IoU (%)		Accuracy (%)		IoU (%)	
		Train	Valid	Train	Valid	Train	Valid	Train	Valid
Cross-Entropy	–	99.82±0.26	99.40±0.07	78.37±1.12	78.37±1.14	99.92±0.21	**94.41±1.07**	80.04±0.66	80.16±0.48
CE w/ Arch	–	99.99±0.02	99.4±0.22	38.58±16.95	39.07±16.98	100±0	95.3±0.3	59.86±17.09	60.6±17.2
CFCE	GT	99.88±0.10	99.32±0.25	83.22±1.13	82.92±1.18	99.96±0.03	92.96±0.15	87.93±0.24	88.16±0.33
CFCE + KL	GT	99.71±0.27	99.32±0.15	**94.93±0.88**	**92.72±0.73**	99.74±0.13	90.08±1.47	**96.22±3.58**	**93.12±2.22**
CFCE	SAM	99.92±0.06	99.37±0.15	83.96±2.1	83.95±2.33	99.6±0.19	93.26±0.67	84.79±1.26	85.26±1.22
CFCE + KL	SAM	99.88±0.07	99.19±0.24	83.46±1.73	83.54±1.96	99.6±0.2	93.7±0.28	84.67±1.16	85.16±1.2
CFCE	BBOX	**100±0.01**	**99.42±0.22**	79.09±2.26	79.13±2.28	**99.98±0**	93.83±0.33	84.26±1.86	84.61±1.91

Notably, KL regularization must not be applied when bounding boxes are used in place of masks, as fitting to the shape of the box mischaracterizes the target. Also note that ground-truth (GT) masks are used for validation in every setting to ensure a fair comparison.

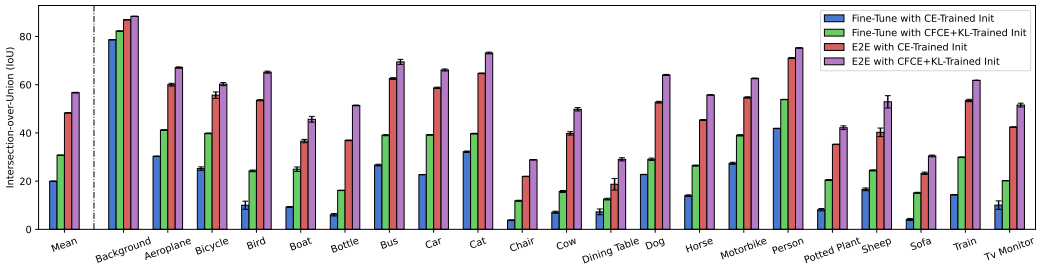
5.3 Semantic Boundaries Dataset (PASCAL VOC)

The Semantic Boundaries Dataset introduces segmentation annotations to the entire Pascal VOC 2011 Dataset (Everingham et al., 2011). We use this dataset to demonstrate performance improvements for both classification and downstream detection settings.

Classification. PASCAL VOC encodes a 20-class *multilabel* classification task; thus input image may contain multiple positive classifications. We report a pareto improvement with increased Average Precision (AP) and Intersection-over-Union (IoU) scores when using core-focused loss formulations.

Method	AP (%)		IoU (%)	
	Train	Valid	Train	Valid
Cross-Entropy	**99.75 ±0.30**	87.32 ±2.58	46.08 ±16.54	44.50 ±16.57
CE w/ Arch	99.57±0.74	**88.85±0.79**	40.69±16.37	38.55±16.43
CFBCE	98.38 ±2.49	88.39 ±1.23	85.00 ±1.32	82.07 ±0.91
CFBCE + KL	97.92 ±1.00	87.19 ±0.46	**89.53 ±1.89**	**85.39 ±0.60**

Segmentation. We also report improvements in IoU performance of core-focused backbones on downstream segmentation, both when fine-tuned (i.e., with a frozen backbone) and trained end-to-end.



6 Discussion

In this work, we establish a connection between interpretability and feature alignment. We demonstrate the impact of utilizing *post-hoc* (i.e., post-training) explainability methods, primarily used as sanity checks, as a guiding factor during training to improve feature alignment with encouraging effect. Core-Focused Cross Entropy is a direct result of the desirable theoretical properties of ContrastiveCAMs, establishing the value of correctness guarantees in interpretability. Reductive metrics inevitably present a partial view of factors that influence model prediction, and comprehensively ensuring that deep neural networks faithfully learn to solve the intended, underlying objective remains a significant challenge for the research community. We hope that our work motivates further exploration towards connections between interpretability and alignment of deep neural networks.

REFERENCES

Ananthu Aniraj, Cassio F Dantas, Dino Ienco, and Diego Marcos. Masking strategies for background bias removal in computer vision models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4397–4405, 2023.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

Matteo Brosolo, P Vinod, and Mauro Conti. Through the static: Demystifying malware visualization via explainability. *Journal of Information Security and Applications*, 91:104063, 2025.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018. doi: 10.1109/wacv.2018.00097. URL http://dx.doi.org/10.1109/WACV.2018.00097.

Adrian Chavarro, Diego Renza, and Ernesto Moya-Albor. Convnext as a basis for interpretability in coffee leaf rust classification. *Mathematics (2227-7390)*, 12(17), 2024.

Yu-Hsi Chen and Chin-Tien Wu. Reynoldsflow: Exquisite flow estimation via reynolds transport theorem, 2025. URL https://arxiv.org/abs/2503.04500.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

Rachel Lea Draelos and Lawrence Carin. Explainable multiple abnormality classification of chest ct volumes. *Artificial Intelligence in Medicine*, 132:102372, 2022.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html, 2011.

Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7):1–39, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.

Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Danny Panknin, and Ahcene Boubekki. Position: Xai needs formal notions of explanation correctness. In *Interpretable AI: Past, Present and Future*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34: 26726–26739, 2021.

Kamal Kc, Zhendong Yin, Dasen Li, and Zhilu Wu. Impacts of background removal on convolutional neural networks for plant disease classification in-situ. *Agriculture*, 11(9):827, 2021.

Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues, June 2022.

Jamuna S Murthy and GM Siddesh. Ai based criminal detection and recognition system for public safety and security using novel criminalnet-228. In *International Conference on Frontiers in Computing and Systems*, pages 3–20. Springer, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

César G Pachón, Dora M Ballesteros, and Diego Renza. An efficient deep learning model using network pruning for fake banknote recognition. *Expert Systems with Applications*, 233:120961, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=XVPqLyNxSyh.

Sahil Singla, Mazda Moayeri, and Soheil Feizi. Core risk minimization using salient imagenet. *arXiv preprint arXiv:2203.15566*, 2022.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

Zifan Wang. *On the Feature Alignment of Deep Vision Models Explainability and Robustness Connected at Hip*. PhD thesis, Carnegie Mellon University, 2023.

Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Information Fusion*, 92:154–176, 2023.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Mathematical Derivations

Proposition 3.1. *The softmax function is invariant to a universal shift of all its input components:*

$$\sigma(\mathbf{x}) = \sigma(\mathbf{x} + a\mathbf{1}_C) \quad \forall \mathbf{x} \in \mathbb{R}^C,\ a \in \mathbb{R}$$

Proof.

$$\sigma(\mathbf{x}+a\mathbf{1}_C) = \frac{1}{\sum_{c=1}^{C} e^{\mathbf{x}_c+a}}(e^{\mathbf{x}_1+a}, e^{\mathbf{x}_2+a}, \cdots, e^{\mathbf{x}_C+a}) = \frac{1}{\sum_{c=1}^{C} e^{\mathbf{x}_c}}(e^{\mathbf{x}_1}, e^{\mathbf{x}_2}, \cdots, e^{\mathbf{x}_C}) = \sigma(\mathbf{x}). \tag{19}$$

□

Theorem 3.2. *HiResCAM explanations* $\mathbf{CAM}^{\text{HiRes}} \in \mathbb{R}^{C\times d_1\times d_2}$ *corresponding to probability predictions* $\tilde{f}(\mathbf{X}) \in \mathbb{R}^C$ *are not uniquely determined, admitting a universal shift of class-level explanations* $\mathbf{CAM}_c^{\text{HiRes}}$ *by an arbitrary matrix* $M \in \mathbb{R}^{d_1\times d_2}$ $\forall c \in [C]$.

$$\tilde{f}(\mathbf{X}) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}^{\text{HiRes}}_{:,i,j} + \mathbf{b}\right) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \overline{\mathbf{CAM}}^{\text{HiRes}}_{:,i,j} + \mathbf{b}\right) \quad \forall M \in \mathbb{R}^{d_1\times d_2} \tag{20}$$

Where $\overline{\mathbf{CAM}}^{\text{HiRes}}$ *is defined as:*

$$\overline{\mathbf{CAM}}^{\text{HiRes}}_c := \mathbf{CAM}^{\text{HiRes}}_c + M \qquad \forall c \in [C] \tag{21}$$

Proof. First, we define the set of all valid shifts $\mathcal{M}$:

$$\text{Let } \mathcal{M} := \{\mathbf{M} \in \mathbb{R}^{C \times d_1 \times d_2} : \mathbf{M}_i = \mathbf{M}_j \quad \forall i, j \in [C]\} \tag{22}$$

The matrix $\mathbf{M}_i \in \mathbb{R}^{d_1 \times d_2}$ can be arbitrary, provided it is constant $\forall i \in [C]$. Thus $|\mathcal{M}| = \infty$. We will show that all HiResCAM explanations that differ by $\mathbf{M} \in \mathcal{M}$ form an equivalence class under the *softmax* operation. Consider the following set:

$$[\overline{\mathbf{CAM}}^{\text{HiRes}}] = \{\mathbf{CAM}^{\text{HiRes}} + \mathbf{M} : \mathbf{M} \in \mathcal{M}\} \tag{23}$$

We then show that any $\overline{\mathbf{CAM}}^{\text{HiRes}}$ with a corresponding shift $\mathbf{M}'$ is a valid explanation (i.e., preserves the final prediction). With logits f deconstructed into HiResCAMs following Eq. (3), we have:

$$\sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \overline{\mathbf{CAM}}^{\text{HiRes}}_{:,i,j} + \mathbf{b}\right) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \left(\overline{\mathbf{CAM}}^{\text{HiRes}}_{:,i,j} + \mathbf{M}'_{:,i,j}\right) + \mathbf{b}\right) \tag{24}$$

Let $a = \sum_{i=1,j=1}^{d_1,d_2} \mathbf{M}'_{c,i,j}$ for some $c \in [C]$. By property of $\mathbf{M}$:

$$= \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \left(\mathbf{CAM}^{\text{HiRes}}_{:,i,j}\right) + a\mathbf{1}_C + \mathbf{b}\right) \tag{25}$$

Applying Proposition 3.1, we have:

$$= \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}^{\text{HiRes}}_{:,i,j} + \mathbf{b}\right) = \tilde{f}(\mathbf{X}) \tag{26}$$

Thus, we have:

$$\tilde{f}(\mathbf{X}) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}^{\text{HiRes}}_{:,i,j} + \mathbf{b}\right) = \sigma\left(\sum_{i=1,j=1}^{d_1,d_2} \overline{\mathbf{CAM}}^{\text{HiRes}}_{:,i,j} + \mathbf{b}\right) \quad \forall M \in \mathbb{R}^{d_1 \times d_2} \tag{27}$$

Proving the desired statement. □

Theorem 3.5. *Let* $\mathbf{CAM}^{\text{HiRes}}$ *and* $\overline{\mathbf{CAM}}^{\text{HiRes}}$ *be two HiResCAMs corresponding to probability predictions* $\tilde{f}(\mathbf{X}) \in \mathbb{R}^C$ *such that:*

$$\overline{\mathbf{CAM}}^{\text{HiRes}} = \left\{\mathbf{CAM}^{\text{HiRes}}_c + M : c \in [C]\right\} \tag{28}$$

Then, for every $M \in \mathbb{R}^{d_1 \times d_2}$*, it holds that:*

$$\mathbf{CAM}^{\text{Cntrst}}_{c_t} = \overline{\mathbf{CAM}}^{\text{Cntrst}}_{c_t} \quad \text{and} \quad \mathbf{CAM}^{\text{Recon}} = \overline{\mathbf{CAM}}^{\text{Recon}} \tag{29}$$

Proof. For some $c_t \in [C]$, we have:

$$\overline{\mathbf{CAM}}^{\text{Cntrst}}_{c_t} = \left\{\overline{\mathbf{CAM}}^{\text{Cntrst}}_{(c_t,c)} : c \in [C] \setminus c_t\right\} \tag{30}$$

$$= \left\{\overline{\mathbf{CAM}}^{\text{HiRes}}_{c_t} - \overline{\mathbf{CAM}}^{\text{HiRes}}_{c} : c \in [C] \setminus c_t\right\} \tag{31}$$

By definition of $\overline{\mathbf{CAM}}^{\text{HiRes}}$, we have:

$$= \left\{\mathbf{CAM}^{\text{HiRes}}_{c_t} + M - \mathbf{CAM}^{\text{HiRes}}_{c} - M : c \in [C] \setminus c_t\right\} \tag{32}$$

$$= \left\{\mathbf{CAM}^{\text{HiRes}}_{c_t} - \mathbf{CAM}^{\text{HiRes}}_{c} : c \in [C] \setminus c_t\right\} \tag{33}$$

$$= \left\{\mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} : c \in [C] \setminus c_t\right\} = \mathbf{CAM}^{\text{Cntrst}}_{c_t} \tag{34}$$

$$\therefore \mathbf{CAM}_{c_t}^{\text{Cntrst}} = \overline{\mathbf{CAM}}_{c_t}^{\text{Cntrst}} \tag{35}$$

This proves the first statement. Now, we can tend to the $\mathbf{CAM}^{\text{Recon}}$ case:

$$\overline{\mathbf{CAM}}_{c_t}^{\text{Recon}} = \overline{\mathbf{CAM}}_{c_t}^{\text{HiRes}} - \frac{1}{C}\sum_{c=1}^{C} \overline{\mathbf{CAM}}_{c}^{\text{HiRes}} \tag{36}$$

By definition of $\overline{\mathbf{CAM}}^{\text{HiRes}}$, we have:

$$= \mathbf{CAM}_{c_t}^{\text{HiRes}} + M - \frac{1}{C}\sum_{c=1}^{C}\left(\mathbf{CAM}_{c}^{\text{HiRes}} + M\right) \tag{37}$$

$$= \mathbf{CAM}_{c_t}^{\text{HiRes}} + M - \frac{C \cdot M}{C} - \frac{1}{C}\sum_{c=1}^{C} \mathbf{CAM}_{c}^{\text{HiRes}} \tag{38}$$

$$= \mathbf{CAM}_{c_t}^{\text{HiRes}} - \frac{1}{C}\sum_{c=1}^{C} \mathbf{CAM}_{c}^{\text{HiRes}} = \mathbf{CAM}_{c_t}^{\text{Recon}} \tag{39}$$

$$\therefore \mathbf{CAM}^{\text{Recon}} = \overline{\mathbf{CAM}}^{\text{Recon}} \tag{40}$$

Proving the desired statements. □

Proposition 4.1. *Softmax-activated class probabilities $\tilde{f}$ can be expressed as a direct function of ContrastiveCAMs and the bias vector.*

$$\tilde{f}_{c_t}(\mathbf{X}) = \left(\sum_{c=1}^{C} \exp\left(\mathbf{b}_c - \mathbf{b}_{c_t} - \sum \mathbf{CAM}_{(c_t,c)}^{\text{Cntrst}}\right)\right)^{-1} \qquad \forall c_t \in [C] \tag{41}$$

Where $\mathbf{CAM}_{(c_t,c_t)}^{\text{Cntrst}} = \mathbf{0}_{d_1 \times d_2}$.

Proof. Individual class probabilities for logit vector f are defined as:

$$\tilde{f}_{c_t} = \sigma_{c_t}(f) = \frac{e^{f_{c_t}}}{\sum_i e^{f_i}} \tag{42}$$

For some $c_t \in [C]$.

We define our logit vector in terms of the elementwise difference to a target class c:

$$\mathbf{d} := f - f_{c_t} \implies f = f_{c_t} + \mathbf{d} \tag{43}$$

Based on this definition, class probabilities can equivalently be computed as:

$$\tilde{f}_{c_t} = \frac{e^{f_{c_t}}}{\sum_i e^{f_i}} = \frac{e^{f_{c_t}}}{\sum_i e^{f_{c_t}+\mathbf{d}_i}} = \frac{e^{f_{c_t}}}{e^{f_{c_t}}\sum_i e^{\mathbf{d}_i}} = \frac{1}{\sum_i e^{\mathbf{d}_i}} \tag{44}$$

This re-contextualizes softmax as a direct function of the differences of class logits. We can further deconstruct the difference by logit values:

$$\mathbf{d}_c = f_c - f_{c_t} = \sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}_{c,i,j}^{\text{HiRes}} + \mathbf{b}_c - \sum_{i=1,j=1}^{d_1,d_2} \mathbf{CAM}_{c_t,i,j}^{\text{HiRes}} - \mathbf{b}_{c_t} \tag{45}$$

Applying Definition 3.3, we have:

$$\mathbf{d}_c = \mathbf{b}_c - \mathbf{b}_{c_t} - \sum \mathbf{CAM}_{(c_t,c)}^{\text{Cntrst}} \tag{46}$$

Substituting $\mathbf{d}_i$ from Eq. (46) into Eq. (44), we have:

$$\tilde{f}_{c_t}(\mathbf{X}) = \frac{1}{\sum_{i=1}^{C} \exp\left(\mathbf{b}_c - \mathbf{b}_{c_t} - \sum \mathbf{CAM}_{(c_t,c)}^{\text{Cntrst}}\right)} = \left(\sum_{c=1}^{C} \exp\left(\mathbf{b}_c - \mathbf{b}_{c_t} - \sum \mathbf{CAM}_{(c_t,c)}^{\text{Cntrst}}\right)\right)^{-1} \tag{47}$$

We can thus compute class probabilities as a direct function of ContrastiveCAMs and the bias vector. □

Proposition 4.2. *Given bias-free classifier h, we can precisely associate the impact of specific regions, encoded by binary mask H, to the computation of cross-entropy loss.*

$$\mathcal{L}_{\text{CE}}(f(\mathbf{X}), \mathbf{y}, H) = \log\left(\sum_{c=1}^{C} \exp\left(-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} - \sum (1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)\right) \tag{48}$$

Proof. Setting $\mathbf{b} = 0$ to the result from Proposition 4.1, we have:

$$\tilde{f}_{c_t}(\mathbf{X}) = \left(\sum_{c=1}^{C} \exp\left(-\sum \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)\right)^{-1} \tag{49}$$

For target class $c_t \in [C]$. Let H and $(1-H)$ define core and non-core masks respectively; these are disjoint. We can use this to further disassociate ContrastiveCAMs:

$$\tilde{f}_{c_t} = \left(\sum_{c=1}^{C} \exp\left(-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} - \sum (1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)\right)^{-1} \tag{50}$$

For one-hot encoded target vector $\mathbf{y}$ and target class index c_t, cross-entropy loss is defined as:

$$\mathcal{L}_{CE}(f(\mathbf{X}), \mathbf{y}, H) = -\sum_{c=1}^{C} \mathbf{y}_c \log \tilde{f}_c = -\log \tilde{f}_{c_t} \tag{51}$$

To which we can substitute softmax using Eq. (50):

$$\begin{aligned}\mathcal{L}_{CE}(f(\mathbf{X}), \mathbf{y}, H) &= -\log\left(\sum_{c=1}^{C} \exp\left(-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} - \sum (1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)\right)^{-1} \\ &= \log\left(\sum_{c=1}^{C} \exp\left(-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} - \sum (1-H) \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)\right)\end{aligned} \tag{52}$$

As core and non-core masks are disjoint, Eq. (52) enables us to identify the logit contributions from the core and non-core regions respectively. □

Theorem 4.6. *A sequence of predictors $f_n \subset \mathcal{F}$ that converges to the optimal $\mathcal{R}_{\text{CFCE}}$-risk also converges to the Bayes-optimal $\mathcal{R}_{\text{CCRM}}$-risk. Equivalently, in the realizable setting, $\mathcal{L}_{\text{CFCE}}$ is classification-calibrated.*

$$\mathcal{R}_{\text{CFCE}}(f_n) \to \mathcal{R}^*_{\text{CFCE}} \implies \mathcal{R}_{\text{CCRM}}(f_n) \to \mathcal{R}^*_{\text{CCRM}} \tag{53}$$

Where $\mathcal{R}_{\text{CFCE}}(f)$ is:

$$\mathcal{R}_{\text{CFCE}}(f) := \mathbb{E}_{(\mathbf{X},(H,\mathbf{y}))\sim\mathcal{D}}\left[\mathcal{L}_{\text{CFCE}}(f(\mathbf{X}), \mathbf{y}, H)\right] \tag{54}$$

Proof. We start by restating Definition (4.5):

$$\begin{aligned}\mathcal{L}_{\text{CFCE}}(f(\mathbf{X}), \mathbf{y}, H)) = \log\Bigg(\sum_{c=1}^{C} \exp\big(&-\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)} \\ &+ \sum (1-H) \odot |\mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}|\big)\Bigg)\end{aligned} \tag{55}$$

$$= \log\left(\sum_{c=1}^{C} \frac{\exp\left(\sum (1-H) \odot |\mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}|\right)}{\exp\left(\sum H \odot \mathbf{CAM}^{\text{Cntrst}}_{(c_t,c)}\right)}\right) \tag{56}$$

We can observe that $\mathcal{R}_{\mathrm{CFCE}}(f)$ takes the following form:

$$\mathcal{R}_{\mathrm{CFCE}}(f) = \mathbb{E}_{(\mathbf{X},(H,\mathbf{y})\sim\mathcal{D}}\left[\log\left(\sum_{c=1}^{C}\underbrace{\frac{\exp\left(\sum(1-H)\odot|\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}|\right)}{\exp\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}\right)}}_{s_c}\right)\right] \tag{57}$$

$\mathcal{R}^*_{\mathrm{CFCE}} = \inf_f \mathcal{R}_{\mathrm{CFCE}}(f)$ is predicated on each summand $s_c \to 0$. We have that:

$$\inf_f\left(\sum_{c=1}^{C}\frac{\exp\left(\sum(1-H)\odot|\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}|\right)}{\exp\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}\right)}\right) \geq \sum_{c=1}^{C}\frac{\inf_f\left(\exp\left(\sum(1-H)\odot|\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}|\right)\right)}{\sup_f\left(\exp\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}\right)\right)} \tag{58}$$

Given sufficiently expressive $\mathcal{F}$ by assumption of realizability of $\mathcal{R}^*_{\mathrm{CCRM}}$, as $n\to\infty,\ f_n$ converges uniformly towards the equality case thus admitting the following dual objective for each s_c:

$$\mathcal{R}_{\mathrm{CFCE}}(f_n)\to\mathcal{R}^*_{\mathrm{CFCE}} \iff \frac{\inf_f\left(\exp\left(\sum(1-H)\odot|\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}|\right)\right)}{\sup_f\left(\exp\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}\right)\right)} \quad \forall c\in[C] \tag{59}$$

With the absolute $|\cdot|$ operator over numerator's exponent and the realizability assumption, we have:

$$\inf_f\left(\exp\left(\sum(1-H)\odot|\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}|\right)\right)=1 \iff \|(1-H)\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c')}\|=0 \tag{60}$$

This satisfies the constraint from Definition 4.4 and further implies (by absolute homogeneity of the norm) that each non-core region has no contribution to the final classification.

Next, we can tend to the denominator.

$$\text{Let } f^* = \arg\sup_f\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}\right) \tag{61}$$

By convexity of $\exp$, we have that:

$$\exp\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst},f^*}_{(c_t,c)}\right)\geq\sup_f\left(\exp\left(\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}\right)\right) \tag{62}$$

The realization of f^* satisfies the following condition:

$$\sum H\odot\mathbf{CAM}^{\mathrm{Cntrst}}_{(c_t,c)}>0 \qquad \forall c\in[C] \tag{63}$$

Which is sufficient to show the largest logit is that of the target class c_t. Thus $\arg\max(f(\mathbf{X})) = \arg\max(\mathbf{y})\ \forall(\mathbf{X},(H,\mathbf{y}))\sim\mathcal{D} \implies \mathbb{E}_{(\mathbf{X},(H,\mathbf{y}))\sim\mathcal{D}}[\ell(f(\mathbf{X},\mathbf{y}))]=0$ which gives us:

$$\mathcal{R}_{\mathrm{CFCE}}(f_n)\to\mathcal{R}^*_{\mathrm{CFCE}} \implies \mathcal{R}_{\mathrm{CCRM}}(f_n)\to\mathcal{R}^*_{\mathrm{CCRM}} \tag{64}$$

Proving the consistency of $\mathcal{L}_{\mathrm{CFCE}}$ as a surrogate minimizer to $\mathcal{R}_{\mathrm{CCRM}}$. □

Proposition B.1. *We can integrate background suppression to the definition of binary cross-entropy using the following formulation:*

$$\mathcal{L}_{\mathrm{CFBCE}}(f(\mathbf{X}),\mathbf{y},H) = -\frac{1}{C}\sum_{i=1}^{C}\Bigg[\mathbf{y}_i\log\left(\phi\left(\sum_{j,k}H_i\odot\mathbf{CAM}^{\mathrm{HiRes}}_{i,j,k}-\sum_{j,k}(1-H_i)\odot|\mathbf{CAM}^{\mathrm{HiRes}}_{i,j,k}|\right)\right) + (1-\mathbf{y}_i)\log\left(1-\tilde{f}(\mathbf{X})_i\right)\Bigg] \tag{65}$$

Proof. We will prove for the multilabel setting, which is a generalization of binary cross-entropy. For binary vector $\mathbf{y}$ (i.e., $\mathbf{y}_i \in \{0, 1\}\ \forall i$), class-specific core masks H_i, and sigmoid ϕ activated logits f, denoted $\tilde{f}$, binary cross-entropy is defined as:

$$\mathcal{L}_{\text{BCE}}(f(\mathbf{X}), \mathbf{y}, H) = -\frac{1}{C}\sum_{i=1}^{C}\left[\mathbf{y}_i \log \tilde{f}_i + (1-\mathbf{y}_i)\log\left(1-\tilde{f}_i\right)\right] \tag{66}$$

$$= -\frac{1}{C}\sum_{i=1}^{C}\left[\mathbf{y}_i \log \phi(f_i) + (1-\mathbf{y}_i)\log\left(1-\phi(f_i)\right)\right] \tag{67}$$

Setting $\mathbf{b} = 0$, we can substitute Eq. (3) within the first term:

$$= -\frac{1}{C}\sum_{i=1}^{C}\left[\mathbf{y}_i \log \phi\left(\sum_{j,k}\mathbf{CAM}^{\text{HiRes}}_{i,j,k}\right) + (1-\mathbf{y}_i)\log\left(1-\tilde{f}_i\right)\right] \tag{68}$$

Similar to Proposition 4.5, we can break down each HiResCAM to core and spurious components. For non-target indices, we seek to reducing logit values across the entire input image. Therefore, we do not disassociate logit values for the second term.

$$\mathcal{L}_{\text{BCE}}(f(\mathbf{X}), \mathbf{y}, H) = -\frac{1}{C}\sum_{i=1}^{C}\left[\mathbf{y}_i \log \phi\left(\sum_{j,k} H_i \odot \mathbf{CAM}^{\text{HiRes}}_{i,j,k} + \sum_{j,k}(1-H_i)\odot \mathbf{CAM}^{\text{HiRes}}_{i,j,k}\right) + (1-\mathbf{y}_i)\log\left(1-\tilde{f}_i\right)\right] \tag{69}$$

The current formulation motivates activating either the core or non-core for positive classification, and motivates de-activating every pixel of the non-positive class. We penalize activation on the non-core regions for the positive class only:

$$\mathcal{L}_{\text{CFBCE}}(f(\mathbf{X}), \mathbf{y}, H) = -\frac{1}{C}\sum_{i=1}^{C}\left[\mathbf{y}_i \log\left(\phi\left(\sum_{j,k} H_i \odot \mathbf{CAM}^{\text{HiRes}}_{i,j,k} - \sum_{j,k}(1-H_i)\odot \left|\mathbf{CAM}^{\text{HiRes}}_{i,j,k}\right|\right)\right) + (1-\mathbf{y}_i)\log\left(1-\tilde{f}(\mathbf{X})_i\right)\right] \tag{70}$$

This gives us the core-focused binary cross-entropy formulation. □

B Core-Focused Cross-Entropic Adaptations

B.1 Core-Focused Binary Cross-Entropy

For sigmoid-activated binary / multilabel classification tasks, we leverage similar principles to define core-focused binary cross-entropy. Since we do not have the contrastive process in softmax-activation, this definitions relies only on HiResCAMs. We represent sigmoid activation using ϕ and admit C target-region masks, denoted H_i for each class $i \in [C]$. In addition, instead of one-hot encoding, we now have binary vector $\mathbf{y}$ (i.e., $\mathbf{y}_i \in \{0, 1\}\ \forall i$).

Proposition B.1 (Core-Focused Binary Cross-Entropy)**.** *We can integrate background suppression to the definition of binary cross-entropy using the following formulation:*

$$\mathcal{L}_{\text{CFBCE}}(f(\mathbf{X}), \mathbf{y}, H) = -\frac{1}{C}\sum_{i=1}^{C}\left[\mathbf{y}_i \log\left(\phi\left(\sum_{j,k} H_i \odot \mathbf{CAM}^{\text{HiRes}}_{i,j,k} - \sum_{j,k}(1-H_i)\odot \left|\mathbf{CAM}^{\text{HiRes}}_{i,j,k}\right|\right)\right) + (1-\mathbf{y}_i)\log\left(1-\tilde{f}(\mathbf{X})_i\right)\right] \tag{71}$$

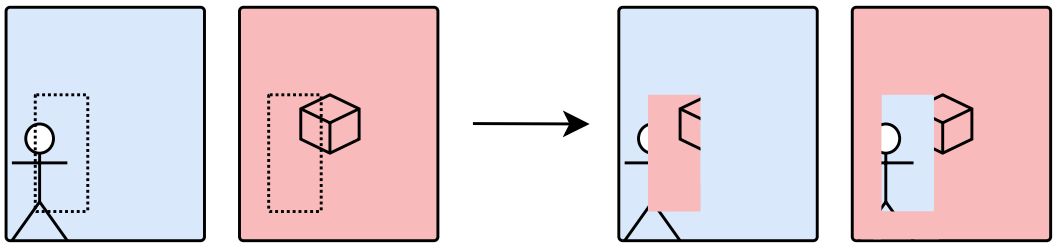
Divergence Regularization Similar to Definition 4.5, we define a divergence term for the target class to motivate activation of the entire core region within the training objective.

Definition B.2 (Regularized Core-Focused Binary Cross-Entropy)**.**

$$\mathcal{L}_{\text{RCFBCE}}(f(\mathbf{X}), \mathbf{y}, H) = \mathcal{L}_{\text{CFBCE}} + \frac{\lambda_1}{\|\mathbf{y}\|_1} \sum_{i=1}^{C} \mathbf{y}_i D_{\text{KL}} \left(\sigma(\lambda_2 H_i) \,||\, \sigma \left(\lambda_3 \mathbf{CAM}_i^{\text{HiRes}} \right) \right) \quad (72)$$

B.2 CutMix with Core-Focused Cross-Entropy

CutMix (Yun et al., 2019) is a batch-wise augmentation technique that encourages better regularization by a) "cutting" out a randomized rectangle (randomized portion remaining consistent across the batch) of a given image and b) "mixing" the cut-out with it's neighbor. The corresponding labels are mixed by a randomly sampled parameter λ.



Definition B.3 (CutMix with Core-Focused Cross-Entropy)**.** Let segmentation mask H take the following form:

$$H := \begin{cases} -1 & \text{pixel does not contain any class} \\ c & \text{pixel contains class } c \end{cases}$$

Also, let $\mathbb{1}_a$ be the indicator function applied elementwise for some $a \in \mathbb{R}$.

Then, Core-Focused Cross Entropy (4.5) with CutMix is formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{CM\_CFBCE}}(f(\mathbf{X}), H, \mathbf{y}) = \log \Bigg(\sum_i \exp \Bigg(& - \sum \mathbb{1}_c(H) \odot \mathbf{CAM}_{(c,i)}^{\text{Cntrst}} \\ & + \sum |\mathbb{1}_{-1}(H) \odot \mathbf{CAM}_{(c,i)}^{\text{Cntrst}}| + \sum \mathbb{1}_i(H) \odot \mathbf{CAM}_{(c,i)}^{\text{Cntrst}} \Bigg) \Bigg) \end{aligned} \quad (73)$$

Where the third newly introduced term within the exponent expresses differential contrast.

C Training Details

Hyperparameters. To *mitigate reward-hacking* our proposed approach, we selected a consistent set of hyperparameters that generally performs well and use it across all our experiments. We train each model using the Adam optimizer (Kingma, 2014) for 150 epochs with a learning rate of $5 \cdot 10^{-4}$, using a linear warmup of 5 epochs followed by Cosine Annealing (Loshchilov and Hutter, 2016) for the remaining 145 epochs. We use a weight decay of 10^{-4}, a batch size of 768. For divergence regularized approaches, we used $\boldsymbol{\lambda} = \{50,\ 10^3,\ 10\}$.

Reproducibility. The source code, datasets, experiments, evals, and model weights are published under a permissive license and can be found at *[redacted for double blind peer-review].*

C.1 Architecture Modifications

The architecture used for training was ResNet-50 (He et al., 2016), initialized with ImageNet. We introduce the following three key modifications:

Removed final downsampling. For images of size $(224, 224)$, the final downsampling layer converts the latent feature embeddings from $d_1 = d_2 = 14$ to 7. This prohibitively reduces the size of the activation map, and making it hard to capture relevant features. We replace the stride of the final downsampling convolution to $(1, 1)$, matching that of the definition used through the rest of ResNet.

Removed final bias. The bias vector $\mathbf{b}$ within h is not involved in the computation of the class activation map. However, it does affect predictions in a way that is not explained by ContrastiveCAMs. To maintain faithfulness of the explanations, we omit the bias from the final model architecture.

Removed final BatchNormalization & ReLU. Since the HiResCAM construction establishes convolution followed immediately by GAP, the standard architecture which uses BatchNormalization & ReLU layers after each convolution, does not directly explain the class score. We therefore neutralize those functions for the final convolutional block. This recovers the faithfulness guarantee.

Note that the above changes correspond only to the final convolutional block of the backbone g and the bias of the linear classifier h; the rest of the architecture remains consistent.