Improving Context-Shift Robustness of Convolutional Networks via Context-Regularized Cross-Entropy

Jinen Setpal
Elmore Family School of Electrical and Computer Engineering
Purdue University
jsetpal@purdue.edu

Chaoyue Liu
Elmore Family School of Electrical and Computer Engineering
Purdue University
cyliu@purdue.edu

Abstract

Image classification models often encounter objects in diverse and unpredictable contexts at test time, which may differ substantially from those seen during training. This phenomenon, termed *context shift*, exposes a key fragility in neural networks: when models rely on non-causal contextual cues for prediction, shifts in context can lead to significant performance degradation. In this paper, we propose a novel training framework to improve the robustness of convolutional neural networks under context shift. Our approach leverages spatial feature attribution to guide models toward making predictions that rely less on contextual regions and more on object-relevant features. As a first step, recognizing a theoretical limitation of existing feature attribution methods we introduce a new variant, *ContrastiveCAMs*, which produces more faithful attribution maps of model predictions. Building on ContrastiveCAMs, we further propose *Context-Regularized Cross-Entropy* (CR-CE), a modification of the standard cross-entropy loss that regularizes the model's attention by suppressing the influence of contextual regions, thereby improving robustness to context shift. We evaluate the effectiveness of our approach on several medium-to-large scale datasets (Waterbirds, Spawrious, ImageNet/ImageNet-BG, Hard-ImageNet), and report consistent improvements in context-shift robustness.

1 Introduction

The vast applications of convolutional neural networks in safety-critical domains such as medical imaging [38, 33], forensic investigation [26] and self-driving [21] make trustworthy and robust predictions essential for large-scale adoption. It has been demonstrated that existing neural networks, despite achieving remarkable performance on large-scale datasets, heavily rely on supporting context (i.e., background area within the images) to determine their predictions [41, 23, 30]. This sensitivity to surrounding context can degrade performance when the same object appears in a different environment [46, 15, 3, 4], whether due to novel object-context combinations in the test data or adversaries manipulating target-agnostic regions to mislead the model.

We highlight this concern in Figure 1 using the Waterbirds-100 dataset [37] which only has the object-context combinations of *waterbird-waterfront* and *landbird-land*. When a *waterbird* is shown in the context of *land*, the model predicts the class *landbird* with high confidence, due to the influence of the context dominating over the influence of the object. The converse also occurs when a *landbird* appears in a *waterfront* context. This concerning lack of robustness has been observed across safety-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

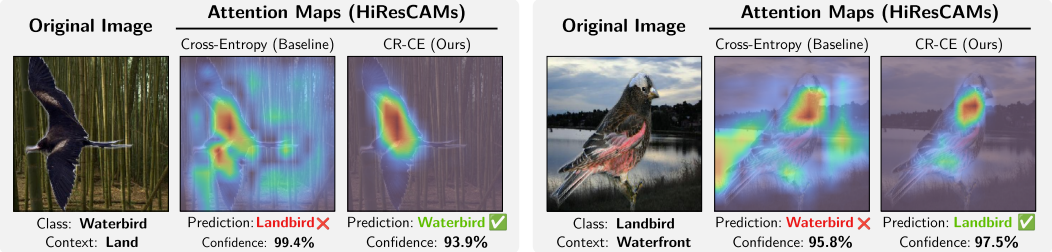


Figure 1: **Context Contributions Can Mislead Predictions.** Spatial feature attribution maps reveal that cross-entropy trained models rely heavily on context, causing misclassifications in novel environments. Our approach encourages focus on the object thereby improving context-shift robustness.

critical applications spanning the diagnosis of skin cancer [27], pneumothorax [36], as well as in the generalization capabilities of robot foundation models [48].

In this paper, we aim to improve the context robustness of convolutional neural networks by leveraging spatial feature attributions. We begin by noting a theoretical limitation of existing feature attribution methods: they permit an arbitrary constant shift applied uniformly across all output logits, which undermines their ability to provide faithful explanations. As a result, the attributed contributions from context and object regions may not accurately reflect their true influence, as they can be distorted by this constant shift. We address this limitation by proposing a new variant *ContrastiveCAMs*, based on a prior method, HiResCAMs [10]. We further theoretically confirm that *ContrastiveCAMs* provide faithful explanations to the neural networks final predictions.

To address sensitivity to context shifts, we propose a new loss function, *Context-Regularized Cross-Entropy* (CR-CE), which leverages the proposed feature attribution method, *ContrastiveCAMs*, to encourage the model to focus on the primary object by suppressing contributions from the surrounding context. This approach requires object-context segmentation during training (crucially, with no segmentation required during inference), which can be readily obtained using existing pretrained segmentation models such as SAM3 [5] and Masked AutoEncoders (MAE) [13, 18]. Notably, the segmentation need not be perfect. In our experiments on the ImageNet dataset, we use MAEs to generate object-context segmentation. Experimental results demonstrate that our method effectively reduces the context contribution to model predictions and improves convolutional networks' robustness to context shifts consistently across a broad class of datasets. Our key contributions are summarized as follows:

- We identify a theoretical limitation of existing spatial feature attribution methods that hinders explanation faithfulness. We address this limitation by proposing a new method *ContrastiveCAMs*.
- We propose a new loss function, *Context-Regularized Cross Entropy* (CR-CE), that encourages model predictions to focus on the object region rather than the context.
- Experimental evaluations demonstrate that the proposed training approach effectively attracts the focus of feature attribution towards the object region, thus enhancing context-shift robustness.

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 $[K] := \{1, 2, \ldots, K\}$.

1.1 Related Work

Feature Attribution Methods in Convolutional Networks. Class Activation Mappings (CAMs) [49] is a prominent family of methods that identify attributions of each (spatial) feature, including GradCAM [39], GradCAM++ [6], Score-CAM [45], and HiResCAM [10]. These methods typically compute feature attributions with respect to the model's output logits, prior to the final softmax activation. However, due to the contrastive nature of the softmax function [28, 31], this can lead to inaccurate attributions for the post-softmax predictions. DiffGradCAM [31] shares a similar approach

to ContrastiveCAMs. But it is based on GradCAM and its connection to the model's prediction is not theoretically clear.

Context-Shift Robustness. One approach to improving robustness is through explanation-guided methods, which leverage feature attribution or saliency maps [40] to direct the model's attention toward relevant objects and away from contextual distractions in the images [14]. This is typically achieved by augmenting the standard cross-entropy loss with additional regularization terms that penalize discrepancies between the model's feature attribution map and a target explanation [35, 29, 34, 11]. The regularization term of Right for the Right Reasons (RRR) [35] penalizes the gradients of log-prediction w.r.t. the irrelevant input features. GALS [29] regularizes the Grad-CAM of the model's predictions toward a target feature attribution map generated using training labels and auxiliary vision-language models. The regularization term of EBL [34] penalizes the positive attributions in contextual regions of the feature attribution map. ABNs [11] introduce an auxiliary attention module to find key regions of images to assist convolutional networks in making predictions. In contrast, our method directly modifies the cross-entropy loss rather than introducing an additional regularization term, so that robustness to context shifts is built in intrinsically.

An orthogonal approach is worst-group generalization [2, 7, 37]. Rather than optimizing for overall accuracy, this line of work partitions the dataset into groups defined by object-context combinations and seeks to balance performance across groups by improving the worst-performing group.

2 Motivation

In image classification tasks, deep neural networks make their predictions based on both the *object* and the *context*, i.e., the backgrounds in images. It has been observed that irrelevant, or even inversely correlated, context can make a substantial contribution to the model predictions, lowering model's confidence on correct predictions, or even overturning the prediction into incorrectness with high confidence [47, 15, 30]. This reliance renders the model less robust to context shifts, such as when a trained model is deployed in a different environment, and more vulnerable to adversarial manipulation; for instance, when an attacker deliberately replaces the context with one commonly associated with other classes. In contrast, humans are able to distinguish between key features and spurious context; where deemed necessary, they can focus on the object region, neglecting the context, so as to accurately classify the relevant object in question.

As an illustration of this issue, consider the Waterbirds dataset [37], which consists of two classes of birds: *waterbirds* and *landbirds*. During training, most birds appear in their natural habitat: waterbirds are present by waterfronts, whereas landbirds are present surrounded by land. However, during testing, a context shift may occur, where a landbird (or a waterbird) appears in a waterfront context (or a land context, respectively). Neural networks often make incorrect predictions on images with context shift. For example, with a fine-tuned ResNet-50, the context region on average accounts for 46% of the model's prediction score compared to the influence from the object. Moreover, the model's accuracy drops from 98% for normal images down to 41% on images with context shift, which is even worse than the random guess accuracy of 50%. Using HiResCAMs [10], a member of the CAM (short for Class Activation Mapping) family that produces a heatmap highlighting the contribution from spatial locations of the image to the network output [49], we can visualize the context's contributions to the model prediction, see Figure 1 for example. We see that when the image context is correlated with a different class, the misleading context makes large contributions to the prediction resulting in misclassifications with high confidence. This shows the pervasiveness of the context robustness concern, supported by broader statistics across other datasets in Table 1.

Our observation is that the neural networks, together with its optimization problem, do not possess an intrinsic mechanism to distinguish objects from context. To see this, consider a neural network f (before the *softmax*), such as a ResNet-50, composed of a convolutional backbone g and a bias-free linear classifier h: $f = h \circ g$. Given an input image $\mathbf{X}$, denote the feature embedding $g(\mathbf{X}) = \mathbf{A} \in \mathbb{R}^{d_0 \times d_1 \times d_2}$ and the final logit $f(\mathbf{X}) \in \mathbb{R}^K$, where K is the number of classes, 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. HiResCAMs, defined as follows,

$$\mathcal{A}_k = \sum_{j=1}^{d_0} (\nabla_{\mathbf{A}_j} f_k) \odot \mathbf{A}_j, \quad \forall k \in [K], \tag{1}$$

have the following relation with the network output logits [10]:

$$f_k(\mathbf{X}) = \sum_{i,j=1}^{d_1,d_2} \left(\mathcal{A}_k(\mathbf{X})\right)_{ij}, \quad \forall k \in [K]. \tag{2}$$

Remark 2.1. Other CAM methods, such as GradCAM [39] and GradCAM++ [6], do not possess the above relation with output logits.

Let $H \in \mathbb{B}^{d_1 \times d_2}$ be the object indicator function, i.e., $H_{ij} = 1$ if the (i,j)-patch corresponds to the object and $H_{ij} = 0$ if the (i,j)-patch corresponds to the context. Then, the network's final output logits can equivalently be written as:

$$f_k(\mathbf{X}) = \sum_{i,j=1}^{d_1,d_2} \left(H \odot \mathcal{A}_k(\mathbf{X})\right)_{ij} + \sum_{i,j=1}^{d_1,d_2} \left((1-H) \odot \mathcal{A}_k(\mathbf{X})\right)_{ij}, \quad \forall k \in [K], \tag{3}$$

where the first and second terms correspond to contributions from the object and context regions respectively. Here, we can see that the context contributes to the prediction in the same way as the object regions, and as seen in the demonstration in Figure 1, the network does not learn to suppress the context contributions.

In this paper, we seek to alleviate this issue of sensitivity to context shift with a focus on convolutional neural networks. Specifically, we will utilize a variant of HiResCAMs to design a new loss function that is capable of suppressing context contributions, encouraging the models to make predictions by primarily focusing on the objects thus improving their robustness to context shift.

3 From HiResCAMs to ContrastiveCAMs

In this section, we first present a fundamental issue of HiResCAMs in the multiclass setting, which degrades the faithfulness of feature attribution maps. We then propose a variant, *ContrastiveCAMs*, which produces accurate, class-specific attribution maps even in the multiclass setting.

HiResCAMs Admit Spurious Shifts. A key drawback of HiResCAMs emerge as a consequence of applying *softmax* activation on the network's output *logits* f to obtain the final predictions $\tilde{f} = \sigma(f)$, which are typically interpreted as *class probabilities*.

Proposition 3.1 (Redundancy in the *softmax* function)**.**

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

Proof. The proof is deferred to Appendix A. □

The above proposition shows that, given a prediction $\tilde{f}$, the logit vector f is not unique: if each logit is shifted by the same value, the final probability predictions $\tilde{f}$ remain unchanged. Therefore, the value of each individual logit f_k alone does not determine the final prediction. This redundancy has also been observed recently by Oh and Noh [28], Piland et al. [31], and discussed in the context of GradCAMs. Here, we show that this redundancy is inherited by HiResCAMs, and further pronounced in terms of degrees of freedom, via Eq. (2).

Theorem 3.2. *Let $\mathcal{A} \in \mathbb{R}^{K \times d_1 \times d_2}$ be a HiResCAM. Given an arbitrary matrix $M \in \mathbb{R}^{d_1 \times d_2}$, define a new HiResCAM as:*

$$\overline{\mathcal{A}}_k := \mathcal{A}_k + M, \qquad \forall k \in [K]. \tag{5}$$

Then, the two HiResCAMs correspond to the same network prediction $\tilde{f}(\mathbf{X})$:

$$\tilde{f}(\mathbf{X}) = \sigma\left(\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{:ij}\right) = \sigma\left(\sum_{i,j=1}^{d_1,d_2} \overline{\mathcal{A}}_{:ij}\right). \tag{6}$$

From this theorem, we can see that the feature attributions rendered by HiResCAMs can be shifted arbitrarily by a matrix $M \in \mathbb{R}^{d_1 \times d_2}$. As a result, feature attributions can spuriously correspond to any

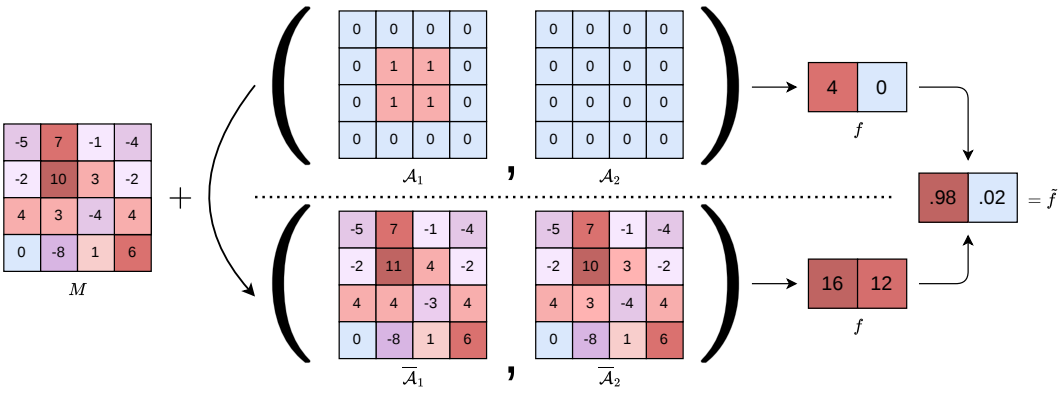


Figure 2: **Illustration of Theorem 3.2.** Shifting $\mathcal{A}$ by arbitrary matrix M results in new explanations $\overline{\mathcal{A}}$. Although the explanation changed significantly, it corresponds to an identical final prediction $\tilde{f}$.

given region without affecting the final prediction. This is illustrated in Figure 2. These additional degrees of freedom hinder the effectiveness of regularizing against object-context segmentation during training, since the explanation can be modified without influencing the overall prediction.

To address this issue, we propose to remove the redundant degrees in the feature attribution maps using a contrastive variant of HiResCAMs, termed *ContrastiveCAMs*.

Definition 3.3 (k-ContrastiveCAMs)**.** Given a class of interest $k \in [K]$ and a set of HiResCAMs $\mathcal{A}$ for a given image $\mathbf{X}$, k-ContrastiveCAMs are defined as follows:

$$\mathcal{C}_k := \left\{\mathcal{C}_{(k,k')} : k' \in [K]\right\}, \qquad \text{where} \qquad \mathcal{C}_{(k,k')} := \mathcal{A}_k - \mathcal{A}_{k'}. \tag{7}$$

The aggregated version of k-ContrastiveCAM is defined as

$$\mathcal{C}_k^{ag} := \frac{1}{K}\sum_{k'=1}^{K} \mathcal{C}_{(k,k')} = \mathcal{A}_k - \frac{1}{K}\sum_{k'=1}^{K} \mathcal{A}_{k'}. \tag{8}$$

Here, $\mathcal{C}_{(k,k')}$ is also referred as (k,k')-ContrastiveCAM. Contrary to HiResCAMs $\mathcal{A}_k$, which focus on single-logit explanations, ContrastiveCAMs provide spatial feature attributions that explain why the model favors one class over a competing class, making them class-discriminative.

Remark 3.4. For binary classification tasks where there is a single output logit and *sigmoid* applied to obtain the final predictions, HiResCAMs are intrinsically contrastive.

Remark 3.5. The idea of subtracting the average CAM has been applied in DiffGradCAMs [31]. However, DiffGradCAMs are built upon GradCAMs, whereas our ContrastiveCAMs are built upon HiResCAMs. With this distinction, both the following key properties hold for ContrastiveCAMs but not for DiffGradCAMs (c.f. Eq.(2) and Remark 2.1).

Proposition 3.6 (Uniqueness of ContrastiveCAMs)**.** *Let $\mathcal{A}$ and $\overline{\mathcal{A}}$ be two HiResCAMs corresponding to probability predictions $\tilde{f}(\mathbf{X}) \in \mathbb{R}^K$ such that:*

$$\overline{\mathcal{A}}_k = \mathcal{A}_k + M, \qquad \forall k \in [K], \tag{9}$$

with $M \in \mathbb{R}^{d_1 \times d_2}$ being an arbitrary matrix. Then, the ContrastiveCAMs $\mathcal{C}$ and $\overline{\mathcal{C}}$, derived from $\mathcal{A}$ and $\overline{\mathcal{A}}$ respectively, are the same:

$$\mathcal{C} = \overline{\mathcal{C}}, \qquad \text{and} \qquad \mathcal{C}^{ag} = \overline{\mathcal{C}}^{ag}. \tag{10}$$

Thus ContrastiveCAMs remove the extra degree of freedom, resulting in faithful feature attribution maps. ContrastiveCAMs also have a close connection to the model's class probabilities. Specifically, any *input-dependent* change to probability predictions $\tilde{f}$ (e.g., as a result of updating model parameters) is reflected precisely by a proportionate change to ContrastiveCAMs.

Proposition 3.7 (Correctness of ContrastiveCAM Explanations)**.** *Given an input* $\mathbf{X}$ *and a class index* $k \in [K]$*, the prediction* $\tilde{f}_k(\mathbf{X})$ *is determined by the* k*-ContrastiveCAMs:*

$$\tilde{f}_k(\mathbf{X}) = \left(\sum_{k'=1}^{K} \exp\left(-\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(k,k')}\right)_{ij} \right) \right)^{-1}, \quad \forall k \in [K]. \tag{11}$$

Moreover, if for some $k \in [K], i \in [d_1], j \in [d_2]$*,* $\left(\mathcal{C}_{(k,k')}\right)_{ij} = 0,\ \forall k' \in [K]$*, then the* (i,j)*-patch does not influence the final prediction.*

Comparing Eq. (11) with Eq. (2), one can easily find that, given any class index k, the corresponding HiResCAM only determines the output logit value f_k but not the final prediction $\tilde{f}_k$; on the other hand, ContrastiveCAMs directly determines the final prediction $\tilde{f}_k$, which makes ContrastiveCAMs a more faithful tool to explain the model's prediction process. Moreover, the above proposition also shows that, iff the ContrastiveCAMs are zero in a certain spatial region, the contribution from this region of the image to the prediction is indeed zero. HiResCAMs do not share this property.

4 Learning with Context-Regularized Cross-Entropy Loss

We now introduce a learning framework designed to be robust to context shifts. In particular, we propose a novel loss function, called *Context-Regularized Cross-Entropy* (CR-CE), which inhibits the context from contributing to the model prediction. This regularization is guided using the feature attributions produced by ContrastiveCAMs.

To effectively regularize against context contributions during training, prior knowledge of the contextual regions in the input images is generally required. Such information can be obtained in two main ways: (i) through segmentation annotations provided by the dataset (e.g., PASCAL VOC [16], MS-COCO [22]), or (ii) by leveraging pretrained object detection or segmentation models (e.g., Segment Anything Model [5], Masked AutoEncoders [18]) to estimate bounding boxes or object masks (see experiments on ImageNet, Spawrious datasets in Section 5). In practice, the first option is often unavailable or prohibitively expensive. We adopt the second approach which is more accessible, as pretrained models can provide approximate context localization with minimal effort. Note that, our framework does not require precise object segmentation or localization, and rough estimates are often sufficient to effectively regularize contextual influence. In Appendix C.4, we show numerical evidence for the robustness of our framework to noisy object-context segmentation.

Given an image $\mathbf{X}$, denote $H \in \mathbb{B}^{d_1 \times d_2}$ as the binary mask that distinguishes object regions from context. Specifically, for each (i,j)-patch where $i \in [d_1], j \in [d_2]$, $H_{ij} = 1$ if the patch corresponds to the object, and $H_{ij} = 0$ if it corresponds to the context. We define CR-CE as follows:

Definition 4.1 (Context-Regularized Cross-Entropy Loss)**.** Given a training dataset $\mathcal{D}$ and convolutional network f, the CR-CE loss is defined as:

$$\mathcal{L}_{\mathrm{CRCE}}(\mathcal{D}) = \mathbb{E}_{(\mathbf{X},\mathbf{y},H)\sim\mathcal{D}}[\ell(\mathbf{X},\mathbf{y},H)], \ \text{ with}$$

$$\ell(\mathbf{X},\mathbf{y},H) = \log\left(\sum_{k=1}^{K} \exp\left(\sum_{i,j=1}^{d_1,d_2} -H_{ij}\left(\mathcal{C}_{(k_t,k)}\right)_{ij} + \underbrace{(1-H_{ij})\left|\left(\mathcal{C}_{(k_t,k)}\right)_{ij}\right|}_{\text{context regularization}} \right) \right). \tag{12}$$

where k_t is the index of the ground-truth class. For simplicity of notation, we do not explicitly write the dependence of H and ContrastiveCAMs $\mathcal{C}$ on the input image $\mathbf{X}$.

The context regularization term in the CR-CE loss regularizes the absolute values of the ContrastiveCAMs $|(\mathcal{C}_{(k_t,k)})_{ij}|$ at context regions, driving them towards zero, thus minimizing the influence of context in the final prediction $\tilde{f}$ (c.f. Proposition 3.7). We also extend context regularization to binary classification settings; we defer this discussion to Appendix B.

Comparison with standard cross-entropy loss. The only difference between the standard cross-entropy (CE) loss and CR-CE lies in the context regularization term. In particular, if the absolute values of the ContrastiveCAM, $|(\mathcal{C}_{(k_t,k)})_{ij}|$, are replaced with their negated values, $-(\mathcal{C}_{(k_t,k)})_{ij}$, the

CR-CE loss reduces to the standard cross-entropy loss, noting that cross-entropy loss on one sample $\ell_{CE}(\mathbf{X}, \mathbf{y}) = -\log \tilde{f}_{k_t}$ as well as the relation in Proposition 3.7.

Despite this seemingly minor modification, the effect on model behavior is substantial. The standard cross-entropy loss implicitly encourages the context to cooperate with the object regions in making predictions. While this is beneficial when the context is consistent with the object, it can lead to an over-reliance on contextual cues and thus reduce robustness under context shift. In contrast, CR-CE explicitly penalizes excessive reliance on context, encouraging the model to base its predictions primarily on the object itself. As a result, when the context is altered or adversarially manipulated, models trained with CR-CE exhibit improved robustness.

Further, for models using a linear classifier head (e.g., ResNet [17], MobileNet [19], EfficientNet [42]), ContrastiveCAM can be computed in closed form, see Eqs. (1) and (7); this results in a minor computational overhead when CR-CE is used in place of cross-entropy.

5 Experimental Results

In this section, we numerically evaluate the effectiveness of training with the proposed CR-CE loss function in both promoting object-focused learning and improving robustness to context shifts.

5.1 Datasets and Experimental Setup

Datasets. We evaluate our method using the following four datasets:

- **Waterbirds-100** dataset [37] consists of images of *waterbirds* and *landbirds*. In the training set, each class of bird appears in the context of their natural habitat–*waterfront* and *land*, respectively. However, in the test set, context shifts may occur, with birds appearing in the opposite environment.
- **Spawrious** dataset [23] consists of AI-generated images of four classes of dog breeds appearing in various contexts. The test set contains combinations of dog breed and context that are not seen during training, thereby assessing models' robustness to context shift.
- **ImageNet** [8] with **ImageNet-BG** [30] as test set. We train the model using ImageNet-1k and test it using ImageNet-BG, the latter of which features images with manually replaced backgrounds, designed to decouple object recognition from context cues.
- **Hard-ImageNet** [25] is a subset of ImageNet, consisting of 15 selected classes of images that have high context influence. It is used to investigate the context's influence on the classification by comparing the model outputs with and without the objects of the images.

For all the datasets above, context shifts only occur in the test set, which is not accessed during training. Ground-truth object-context masks are available for the training sets of Waterbirds-100 and Hard-ImageNet. For Spawrious and ImageNet where object-context segmentation is not available, we use existing segmentation models such as SAM3 [5] and Masked AutoEncoders (MAE) [18, 13] to generate these masks. Our approach allows for imperfect segmentation, enabling the effective use of off-the-shelf vision models when ground truth masks are unavailable.

Experimental Setup. We follow the methodology in the benchmark datasets [37, 23, 30, 25] and fine-tune ResNet-50 models pre-trained on ImageNet-1k. In our experiments, we also evaluate a modified ResNet architecture where the final downsampling convolution layer's stride is set to $(1, 1)$, thus doubling the resolution of ContrastiveCAMs, i.e., $d_1 = d_2 = 14$. The higher-resolution attribution maps provide more precise localization of object regions. More experiments with different architectures can be found in Appendix C.5. We compare our proposed CR-CE framework with the following prior approaches:

- The standard training procedure with cross-entropy loss (CE), which does not implement any context-shift robustness technique. It serves as a baseline for comparison.
- Explanation-guided methods, which improve context-shift robustness by guiding models' feature attributions, i.e., the heatmaps, towards the target object region. Our proposed method falls in this category. We compare with existing methods: attention branch networks (ABN) [11], GALS [29], and training with energy-based loss (EBL) [34].

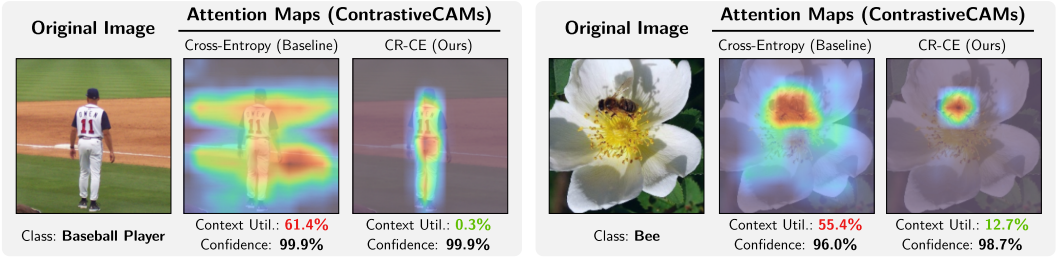


Figure 3: **Qualitative Evaluation.** Models trained using cross-entropy rely heavily on context: in the Hard-ImageNet (left) and ImageNet-BG (right) examples shown above, the *baseball player* and *bee* are predicted using the *playing field* (61.4% of the prediction) and the *flower* (55.4% of the prediction) respectively. On the other hand, models trained using CR-CE leverage object-specific regions and localize the target without the use of object-context segmentation during inference.

Table 1: **Context Contribution Ratios.** On the test set of each context-shift robustness benchmark we consider, CR-CE trained models primarily rely on object-specific regions to determine predictions.

CAM Size	Approach	Context Contribution Ratio $R\%$ ($\downarrow$)			
		Waterbirds-100	Spawrious‡	ImageNet-BG	Hard-ImageNet
7×7	CE (Baseline)	$46.08_{\pm 0.43}$	$44.79_{\pm 0.3}$	$31.83_{\pm 0}$	$68.53_{\pm 0.04}$
	CR-CE (Ours)	$\mathbf{12.86}_{\pm 0.09}$	$\mathbf{14.89}_{\pm 0.2}$	$\mathbf{14.85}_{\pm 0}$	$\mathbf{35.85}_{\pm 0.26}$
14×14	CE (Baseline)	$58.75_{\pm 0.12}$	$47.32_{\pm 0.09}$	$28.27_{\pm 0.27}$	$66.5_{\pm 0.07}$
	CR-CE (Ours)	$\mathbf{9.54}_{\pm 0.28}$	$\mathbf{10.98}_{\pm 0.12}$	$\mathbf{9.32}_{\pm 0.06}$	$\mathbf{31.12}_{\pm 0.1}$

† No std. dev. is reported; we use pre-trained weights from `torchvision` without modification.
‡ We report metrics on the hardest subtask and difficulty (M2M-hard) within the Spawrious benchmark.

- Other methods, including Context Masking training [1] and Causal Feature Alignment [44]. Both methods use object-context masks to transform the input/hidden features in order to discourage context reliance. Particularly, Context Masking only uses the object regions of images for training by masking out all context regions.

For fairness of comparison, we use identical object-context masks across approaches. For all our metrics, we report the mean and standard deviation computed over 3 independent runs. Details of the experimental setup are described in Appendix E.

5.2 CR-CE Suppresses Context Contributions

First, we evaluate the proportion of context contributions using the following metric:

$$R(\mathcal{C}^{ag}, H) = \frac{\sum_{i,j=1}^{d_1,d_2} |(\mathcal{C}^{ag}_{k_t})_{ij}|(1 - H_{ij})}{\sum_{i,j=1}^{d_1,d_2} |(\mathcal{C}^{ag}_{k_t})_{ij}|}, \tag{13}$$

where $k_t \in [K]$ is the target class index. This metric represents the portion of the absolute value of the *ContrastiveCAM* heatmap from the context region. A comparison of the proposed CR-CE training framework with the standard CE training is shown in Table 1, with Figure 3 and Figure 5 presenting examples. Comparisons with the other methods assessed in this work are included in Table 5. We observe that training with CR-CE effectively guides ContrastiveCAMs onto the object regions and reduces the model's reliance on the context region.

Additional evaluation of the effectiveness of CR-CE in suppressing context contributions using the Hard-ImageNet evaluation suite can be found in Appendix C.2.

Table 2: **Context-Shift Robustness on Waterbirds-100, ImageNet, and Spawrious datasets.** Accuracy across various benchmarks, with the best context-shifted performance highlighted in **bold** and the highest in-distribution performance underlined. Models trained with CR-CE generalize effectively to novel environments at a minor cost to in-distribution performance.

CAM Size	Approach	Waterbirds-100		ImageNet/ImageNet-BG		Spawrious‡		
		No Shift (↑)	Shifted (↑)	No Shift (↑)	Shifted (↑)	Easy (↑)	Medium (↑)	Hard (↑)
7×7	CE (Baseline)	$\underline{98.66}_{\pm 0.10}$	$41.74_{\pm 4.15}$	$\underline{82.92}^{\dagger}$	$74.83^{\dagger}$	$90.86_{\pm 0.54}$	$79.46_{\pm 0.48}$	$81.52_{\pm 1.03}$
	GALS	$97.54_{\pm 1.01}$	$54.45_{\pm 6.75}$	$82.68_{\pm 0.06}$	$73.82_{\pm 0.10}$	$93.47_{\pm 0.36}$	$84.58_{\pm 0.88}$	$78.74_{\pm 0.97}$
	EBL	$96.35_{\pm 0.49}$	$51.31_{\pm 4.70}$	$82.88_{\pm 0.08}$	$74.79_{\pm 0.01}$	$92.43_{\pm 1.31}$	$87.23_{\pm 0.84}$	$81.21_{\pm 0.74}$
	CFA	$93.43_{\pm 0.13}$	$72.71_{\pm 0.64}$	$82.71_{\pm 0.03}$	$73.85_{\pm 0.09}$	$89.14_{\pm 0.89}$	$85.68_{\pm 0.68}$	$81.77_{\pm 1.73}$
	Context Masking	$91.16_{\pm 3.18}$	$61.67_{\pm 6.17}$	$80.21_{\pm 0.07}$	$72.87_{\pm 0.10}$	$92.08_{\pm 1.04}$	$85.03_{\pm 2.37}$	$78.46_{\pm 1.12}$
	CR-CE (Ours)	$94.38_{\pm 1.21}$	$\mathbf{73.30}_{\pm 4.43}$	$81.85_{\pm 0.03}$	$\mathbf{75.75}_{\pm 0.08}$	$\mathbf{95.55}_{\pm 0.16}$	$\mathbf{91.25}_{\pm 0.18}$	$\mathbf{87.49}_{\pm 0.14}$
14×14	CE (Baseline)	$\underline{98.48}_{\pm 0.09}$	$38.40_{\pm 1.79}$	$82.33_{\pm 0.10}$	$74.72_{\pm 0.06}$	$88.98_{\pm 0.14}$	$73.38_{\pm 0.87}$	$79.76_{\pm 1.01}$
	ABN	$97.39_{\pm 0.55}$	$51.87_{\pm 6.14}$	$79.91_{\pm 0.06}$	$69.07_{\pm 0.15}$	$89.65_{\pm 0.44}$	$72.81_{\pm 1.32}$	$63.96_{\pm 6.54}$
	GALS	$97.39_{\pm 0.28}$	$63.21_{\pm 4.90}$	$82.43_{\pm 0.04}$	$75.21_{\pm 0.02}$	$95.03_{\pm 1.35}$	$87.63_{\pm 3.80}$	$84.08_{\pm 1.87}$
	EBL	$96.33_{\pm 0.95}$	$51.49_{\pm 5.31}$	$\underline{82.72}_{\pm 0.09}$	$75.79_{\pm 0.20}$	$93.46_{\pm 0.73}$	$89.05_{\pm 0.47}$	$79.36_{\pm 2.83}$
	CFA	$93.09_{\pm 0.74}$	$77.51_{\pm 3.89}$	$79.54_{\pm 0.09}$	$73.48_{\pm 0.03}$	$87.22_{\pm 0.78}$	$85.68_{\pm 0.68}$	$75.88_{\pm 1.12}$
	Context Masking	$92.27_{\pm 1.18}$	$64.39_{\pm 1.79}$	$80.70_{\pm 0.07}$	$75.49_{\pm 0.10}$	$90.96_{\pm 0.24}$	$85.70_{\pm 2.99}$	$80.80_{\pm 0.71}$
	CR-CE (Ours)	$94.36_{\pm 0.12}$	$\mathbf{79.13}_{\pm 0.77}$	$81.25_{\pm 0.03}$	$\mathbf{76.84}_{\pm 0.12}$	$\mathbf{96.47}_{\pm 0.06}$	$\mathbf{92.41}_{\pm 0.29}$	$\mathbf{88.66}_{\pm 0.52}$

† No std. dev. is reported; we use pre-trained weights from `torchvision` without modification.
‡ The reported metrics correspond to the many-to-many spurious correlation setting; the complete benchmark is assessed in Appendix C.1.

5.3 Reduced Context-Contributions Improve Robustness to Context Shifts

Next, we demonstrate the improved performance of CR-CE trained models under context shift by comparing them against the standard cross-entropy training, existing explanation-guided learning approaches [29, 11, 34] and other methods [44, 1]. We report the numerical results in Table 2.

We observe that across all datasets, the CR-CE training framework consistently improves the robustness to context shift and further outperforms models trained using cross-entropy loss and the other compared methods. This demonstrates that the proposed method successfully suppresses contribution from irrelevant or misleading context, yielding correct and robust model predictions. We also note that, when the context is relevant (as in the setting without context shift), there is a slight degradation in the prediction accuracy across tasks. This is expected because when the test dataset and training dataset have the same set of object-context combinations, the model can often infer the correct class with the help of the context. Indeed, such correlations are often not desirable if context-shift robustness is the target.

Additionally, we observe that some existing explanation-guided methods may underperform the standard cross-entropy baseline on complex classification tasks, such as ImageNet, while outperforming it on simpler tasks such as the Waterbirds-100 dataset. We hypothesize that this behavior is largely due to their reliance on classical feature attribution methods (e.g. Grad-CAM and HiResCAM), which suffer from the unfaithfulness issue theoretically described in Theorem 3.2. In the binary classification setting of Waterbirds-100, HiResCAM coincides with our ContrastiveCAMs (see Remark 3.4) and does not exhibit this limitation. More discussion of explanation-guided methods on this matter can be found in Appendix C.3.

Notably, we emphasize that our approach does not use object-context segmentation during inference; the promising performance of our approach suggests that the CR-CE trained models have the ability to distinguish objects from context while making predictions.

Finally, we also note that the generalization gap between in-distribution and context-shifted images shrinks as explanation granularity is increased (comparing the 7×7 and 14×14 CAM resolutions).

6 Conclusion

In this work, we advance explanation faithfulness and context robustness in convolutional networks. We introduce *ContrastiveCAMs*, a feature attribution method that addresses limitations of existing CAM methods and has a strong connection to model predictions. We further propose a new loss function, *Context-Regularized Cross-Entropy* (CR-CE), which penalizes contributions from context regions, thereby improving models' robustness to context shifts. Our experimental evaluations show that our method effectively attracts the feature attributions towards the object region, and enhances context-shift robustness.

Limitations. We identify two limitations of this work. First, the method does not naturally extend to transformer architectures, such as Vision Transformers (ViTs) [9]. This is because CAM-based spatial feature attribution methods rely on the local structure induced by convolutional layers, whereas attention mechanisms [43] mix patch tokens in a way that loses explicit spatial locality. Second, the proposed CR-CE training requires additional object-context segmentation information (albeit only during training), which requires additional cost to obtain in practice. We note that existing explanation-guided methods also rely on this information.

Acknowledgment

This research used both the DeltaAI advanced computing and data resource, which is supported by the National Science Foundation (award OAC 2320345) and the State of Illinois, and the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta and DeltaAI are joint efforts of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

References

[1] 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.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Pedro RAS Bassi, Sergio SJ Dertkigil, and Andrea Cavalli. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications*, 15(1):291, 2024.

[4] Christopher Boland, Keith A Goatman, Sotirios A Tsaftaris, and Sonia Dahdouh. There are no shortcuts to anywhere worth going: Identifying shortcuts in deep learning models for medical image analysis. In *Medical Imaging with Deep Learning*, pages 131–150. PMLR, 2024.

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pages 138846–138923, 2026.

[6] 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`.

[7] Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. In *International Conference on Learning Representations*, 2022.

[8] 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.

[9] 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. In *International Conference on Learning Representations*, 2021.

[10] 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.

[11] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10705–10714, 2019.

[12] Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations. In *International conference on machine learning*, pages 10800–10834. PMLR, 2023.

[13] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7457–7476, 2022.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[20] 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.

[21] 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.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Aengus Lynch, Gbetondji Jean-Sebastien Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

[24] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19087–19097, 2022.

[25] Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *Advances in Neural Information Processing Systems*, 35:10068–10077, 2022.

[26] 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.

[27] Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40, 2021.

[28] Yoojin Oh and Junhyug Noh. Beyond softmax: Dual-branch sigmoid architecture for accurate class activation maps. *arXiv preprint arXiv:2511.05590*, 2025.

[29] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Łukasz Piekarek and Kamil Szyc. Imagenet-bg: A toolkit and dataset for evaluating vision model robustness against background variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 231–240, October 2025.

[31] Jacob Piland, Christopher Sweet, and Adam Czajka. Diffgradcam: A class activation map using the full model decision to solve unaddressed adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2026.

[32] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[33] 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.

[34] Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1922–1933, 2023.

[35] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL `https://doi.org/10.24963/ijcai.2017/371`.

[36] Khaled Saab, Sarah Hooper, Mayee Chen, Michael Zhang, Daniel Rubin, and Christopher Ré. Reducing reliance on spurious features in medical image classification with spatial specificity. In *Machine Learning for Healthcare Conference*, pages 760–784. PMLR, 2022.

[37] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[38] DR Sarvamangala and Raghavendra V Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, 15(1):1–22, 2022.

[39] 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.

[40] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[41] 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`.

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

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] Rahul Venkataramani, Parag Dutta, Vikram Melapudi, and Ambedkar Dukkipati. Causal feature alignment: Learning to ignore spurious background features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4666–4674, 2024.

[45] 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.

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

[47] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[48] Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation. In *Conference on Robot Learning*, pages 3239–3266. PMLR, 2025.

[49] 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.

A Proofs

Proposition 3.1 (Redundancy in the *softmax* function)**.**

$$\sigma(\mathbf{x}) = \sigma(\mathbf{x} + a\mathbf{1}_K), \qquad \forall \mathbf{x} \in \mathbb{R}^K,\ a \in \mathbb{R}. \tag{14}$$

Proof.

$$\sigma(\mathbf{x}+a\mathbf{1}_K) = \frac{1}{\sum_{k=1}^K e^{\mathbf{x}_k+a}}(e^{\mathbf{x}_1+a}, e^{\mathbf{x}_2+a}, \cdots, e^{\mathbf{x}_K+a}) = \frac{1}{\sum_{k=1}^K e^{\mathbf{x}_k}}(e^{\mathbf{x}_1}, e^{\mathbf{x}_2}, \cdots, e^{\mathbf{x}_K}) = \sigma(\mathbf{x}). \tag{15}$$

□

Theorem 3.2. *Let $\mathcal{A} \in \mathbb{R}^{K \times d_1 \times d_2}$ be a HiResCAM. Given an arbitrary matrix $M \in \mathbb{R}^{d_1 \times d_2}$, define a new HiResCAM as:*

$$\overline{\mathcal{A}}_k := \mathcal{A}_k + M, \qquad \forall k \in [K]. \tag{16}$$

Then, the two HiResCAMs correspond to the same network prediction $\tilde{f}(\mathbf{X})$:

$$\tilde{f}(\mathbf{X}) = \sigma\left(\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{:ij}\right) = \sigma\left(\sum_{i,j=1}^{d_1,d_2} \overline{\mathcal{A}}_{:ij}\right). \tag{17}$$

Proof. The proof for this statement follows from the relation between the two HiResCAMs. For a given shift M, the predictions from the corresponding HiResCAM $\overline{\mathcal{A}}$ can be expressed as:

$$\sigma\left(\sum_{i,j=1}^{d_1,d_2} \overline{\mathcal{A}}_{:ij}\right) = \sigma\left(\sum_{i,j=1}^{d_1,d_2} (\mathcal{A}_{:ij} + M_{ij})\right). \tag{18}$$

Let $a = \sum_{i,j=1}^{d_1,d_2} M_{ij}$. We can equivalently write the above statement as:

$$= \sigma\left(\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{:ij} + a\mathbf{1}_K\right). \tag{19}$$

Applying Proposition 3.1, we have:

$$\sigma\left(\sum_{i,j=1}^{d_1,d_2} \overline{\mathcal{A}}_{:ij}\right) = \sigma\left(\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{:ij}\right) = \tilde{f}(\mathbf{X}). \tag{20}$$

Proving the desired statement. □

Proposition 3.6 (Uniqueness of ContrastiveCAMs)**.** *Let $\mathcal{A}$ and $\overline{\mathcal{A}}$ be two HiResCAMs corresponding to probability predictions $\tilde{f}(\mathbf{X}) \in \mathbb{R}^K$ such that:*

$$\overline{\mathcal{A}}_k = \mathcal{A}_k + M, \qquad \forall k \in [K], \tag{21}$$

with $M \in \mathbb{R}^{d_1 \times d_2}$ being an arbitrary matrix. Then, the ContrastiveCAMs $\mathcal{C}$ and $\overline{\mathcal{C}}$, derived from $\mathcal{A}$ and $\overline{\mathcal{A}}$ respectively, are the same:

$$\mathcal{C} = \overline{\mathcal{C}}, \qquad \text{and} \qquad \mathcal{C}^{ag} = \overline{\mathcal{C}}^{ag}. \tag{22}$$

Proof. First, consider k-ContrastiveCAMs. We have that:

$$\begin{aligned}
\overline{\mathcal{C}}_k &= \{\overline{\mathcal{C}}_{(k,k')} : k' \in [K]\} && (23)\\
&= \{\mathcal{A}_k + M - \mathcal{A}_{k'} - M : k' \in [K]\} && (24)\\
&= \{\mathcal{A}_k - \mathcal{A}_{k'} : k' \in [K]\} && (25)\\
&= \{\mathcal{C}_{(k,k')} : k' \in [K]\} = \mathcal{C}_k, \quad \forall k \in [K]. && (26)
\end{aligned}$$

Similarly, in the case of aggregated k-ContrastiveCAMs, we have:

$$\overline{\mathcal{C}}_k^{ag} = \overline{\mathcal{A}}_k - \frac{1}{K}\sum_{k'=1}^{K}\overline{\mathcal{A}}_{k'} \tag{27}$$

$$= \mathcal{A}_k + M - \frac{1}{K}\sum_{k'=1}^{K}(\mathcal{A}_{k'} + M) \tag{28}$$

$$= \mathcal{A}_k + M - \frac{K\cdot M}{K} - \frac{1}{K}\sum_{k'=1}^{K}\mathcal{A}_{k'} \tag{29}$$

$$= \mathcal{A}_k - \frac{1}{K}\sum_{k'=1}^{K}\mathcal{A}_{k'} = \mathcal{C}_k^{ag}, \quad \forall k \in [K]. \tag{30}$$

Eqs. (26) and (30) together prove the desired equalities. □

Proposition 3.7(Correctness of ContrastiveCAM Explanations)**.** *Given an input* $\mathbf{X}$ *and a class index* $k \in [K]$*, the prediction* $\tilde{f}_k(\mathbf{X})$ *is determined by the* k*-ContrastiveCAMs:*

$$\tilde{f}_k(\mathbf{X}) = \left(\sum_{k'=1}^{K}\exp\left(-\sum_{i,j=1}^{d_1,d_2}\left(\mathcal{C}_{(k,k')}\right)_{ij}\right)\right)^{-1}, \quad \forall k \in [K]. \tag{31}$$

Moreover, if for some $k \in [K], i \in [d_1], j \in [d_2]$*,* $\left(\mathcal{C}_{(k,k')}\right)_{ij} = 0,\ \forall k' \in [K]$*, then the* (i,j)*-patch does not influence the final prediction.*

Proof. For a class index $k \in [K]$, the prediction probability given a logit vector f is expressed as:

$$\tilde{f}_k = \sigma_k(f) = \frac{e^{f_k}}{\sum_{k'=1}^{K} e^{f_{k'}}}. \tag{32}$$

We define the logit vector in terms of the elementwise difference to a candidate class k:

$$\mathbf{d} := f - f_k \implies f = f_k + \mathbf{d}. \tag{33}$$

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

$$\tilde{f}_k = \frac{e^{f_k}}{\sum_{k'=1}^{K} e^{f_{k'}}} = \frac{e^{f_k}}{\sum_{k'=1}^{K} e^{f_k+\mathbf{d}_{k'}}} = \frac{e^{f_k}}{e^{f_k}\sum_{k'=1}^{K} e^{\mathbf{d}_{k'}}} = \frac{1}{\sum_{k'=1}^{K} e^{\mathbf{d}_{k'}}}. \tag{34}$$

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

$$\mathbf{d}_{k'} = f_{k'} - f_k = \sum_{i,j=1}^{d_1,d_2}\mathcal{A}_{k'ij} - \sum_{i,j=1}^{d_1,d_2}\mathcal{A}_{kij}, \quad \forall k, k' \in [K]. \tag{35}$$

Applying Definition 3.3, we have:

$$\mathbf{d}_{k'} = \sum_{i,j=1}^{d_1,d_2}(\mathcal{A}_{k'} - \mathcal{A}_k)_{ij} = -\sum_{i,j=1}^{d_1,d_2}\left(\mathcal{C}_{(k,k')}\right)_{ij}, \quad \forall k, k' \in [K]. \tag{36}$$

Substituting $\mathbf{d}_{k'}$ from Eq. (36) into Eq. (34), we have:

$$\tilde{f}_k(\mathbf{X}) = \frac{1}{\sum_{k'=1}^{K}\exp\left(-\sum_{i,j=1}^{d_1,d_2}\left(\mathcal{C}_{(k,k')}\right)_{ij}\right)} \tag{37}$$

$$= \left(\sum_{k'=1}^{K}\exp\left(-\sum_{i,j=1}^{d_1,d_2}\left(\mathcal{C}_{(k,k')}\right)_{ij}\right)\right)^{-1}, \quad \forall k \in [K]. \tag{38}$$

We can thus compute class probabilities as a direct function of ContrastiveCAMs.

Next, we show that zero-patches do not influence predictions. For a given class $k \in [K]$, denote the set of zero-patches as $\mathcal{Z}_k$, and its complement as $\mathcal{Z}'_k$:

$$\mathcal{Z}_k = \left\{(i,j) \in [d_1] \times [d_2] : \left(\mathcal{C}_{(k,k')}\right)_{ij} = 0,\ \forall k' \in [K]\right\}, \quad \text{and} \quad \mathcal{Z}'_k = ([d_1] \times [d_2]) \setminus \mathcal{Z}_k. \tag{39}$$

Applying this partition to Eq. (38), we have that:

$$\tilde{f}_k(\mathbf{X}) = \left(\sum_{k'=1}^{K} \exp\left(-\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(k,k')}\right)_{ij}\right)\right)^{-1} \tag{40}$$

$$= \left(\sum_{k'=1}^{K} \exp\left(-\sum_{i,j\in\mathcal{Z}'_k} \left(\mathcal{C}_{(k,k')}\right)_{ij} - \sum_{i,j\in\mathcal{Z}_k} \left(\mathcal{C}_{(k,k')}\right)_{ij}\right)\right)^{-1} \tag{41}$$

$$= \left(\sum_{k'=1}^{K} \exp\left(-\sum_{i,j\in\mathcal{Z}'_k} \left(\mathcal{C}_{(k,k')}\right)_{ij} - \sum_{i,j\in\mathcal{Z}_k} 0\right)\right)^{-1} \tag{42}$$

$$= \left(\sum_{k'=1}^{K} \exp\left(-\sum_{i,j\in\mathcal{Z}'_k} \left(\mathcal{C}_{(k,k')}\right)_{ij}\right)\right)^{-1}. \tag{43}$$

Note that, for some $k \in [K]$, $\left(\mathcal{C}_{(k,k')}\right)_{ij} = 0$ for all $k' \in [K]$ is equivalent to $(\mathcal{A}_k)_{ij} = (\mathcal{A}_{k'})_{ij}$ for all $k' \in [K]$, which further implies that $\left(\mathcal{C}_{(k,k')}\right)_{ij} = 0$ for all $k, k' \in [K]$. This means that $\mathcal{Z}_k = \mathcal{Z}_{k'}$ for all $k, k' \in [K]$. Therefore, Eq. (43) holds for all $k \in [K]$, implying all the logit outputs are not affected by the patch (i, j). □

Remark 3.4. *For binary classification tasks where there is a single output logit and sigmoid applied to obtain the final predictions, HiResCAMs are intrinsically contrastive.*

Proof. For $K = 2$, i.e., binary classification tasks, one can obtain class probabilities using two setups: (i) $f \in \mathbb{R}^2$ with softmax σ activation, or (ii): $f' \in \mathbb{R}$ with sigmoid ϕ activation.

Consider ContrastiveCAMs $\mathcal{C} \in \mathbb{R}^{2\times d_1 \times d_2}$ for softmax activation, and the corresponding HiResCAMs $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2}$ for sigmoid activation. We will prove that HiResCAMs in the sigmoid setup are equivalent to ContrastiveCAMs in the softmax setup.

For the softmax case, we apply Proposition 3.7 and obtain class probabilities as:

$$p_f(y = 1|\mathbf{X}) = \tilde{f}_1(\mathbf{X}) = \left(\exp\left(-\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(1,1)}\right)_{ij}\right) + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(1,2)}\right) ij\right)\right)^{-1}. \tag{44}$$

$\mathcal{C}_{(k,k)} = \mathbf{0}_{d_1 \times d_2},\ \forall k \in [K]$, hence we have:

$$p_f(y = 1|\mathbf{X}) = \left(1 + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(1,2)}\right)_{ij}\right)\right)^{-1} \tag{45}$$

$$= \left(1 + \exp\left(\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(2,1)}\right)_{ij}\right)\right)^{-1}, \tag{46}$$

and similarly,

$$p_f(y = 2|\mathbf{X}) = \left(1 + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \left(\mathcal{C}_{(2,1)}\right)_{ij}\right)\right)^{-1}. \tag{47}$$

For the sigmoid case, we apply Proposition B.1 and obtain class probabilities as:

$$p_{f'}(y=1|\mathbf{X}) = 1 - \phi(f') = \phi(-f') = \left(1 + \exp\left(\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{ij}\right)\right)^{-1}, \tag{48}$$

and

$$p_{f'}(y=2|\mathbf{X}) = \phi(f') = \left(1 + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{ij}\right)\right)^{-1}. \tag{49}$$

Comparing Eq. (46) with Eq. (48) and Eq. (47) with Eq. (49), we observe that HiResCAMs' connection to the model prediction is identical to that of ContrastiveCAMs, hence HiResCAMs in the binary setting are intrinsically contrastive. □

Proposition B.1 (Correctness of HiResCAMs in Binary Setting)**.** *Given an input* $\mathbf{X}$*, the prediction* $\tilde{f}(\mathbf{X})$ *of a binary classifier is determined by the corresponding HiResCAM:*

$$\tilde{f}(\mathbf{X}) = \left(1 + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{ij}\right)\right)^{-1}. \tag{50}$$

Proof. In the binary classification setting, $f(\mathbf{X}) \in \mathbb{R}$, and $\tilde{f} = \phi(f)$, where ϕ denotes *sigmoid* activation. Applying Eq. (2), we have:

$$\tilde{f}(\mathbf{X}) = \phi(f) = (1 + \exp(-f))^{-1} = \left(1 + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{ij}\right)\right)^{-1}. \tag{51}$$

□

B Context-Regularized Binary Cross-Entropy

In the binary classification setting, the neural network outputs a single logit, $f(\mathbf{X}) \in \mathbb{R}$, to which the *sigmoid* function, denoted $\phi : \mathbb{R} \to [0, 1]$, is applied to produce the corresponding probability predictions: $\tilde{f} = \phi(f)$. The corresponding class labels are denoted $y \in \{0, 1\}$. Additionally, HiResCAMs have shape $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2}$, and their connection to logits in Eq. (2) extends to class probabilities. This follows from the bijectivity of *sigmoid* activation:

Proposition B.1 (Correctness of HiResCAMs in Binary Setting)**.** *Given an input* $\mathbf{X}$*, the prediction* $\tilde{f}(\mathbf{X})$ *of a binary classifier is determined by the corresponding HiResCAM:*

$$\tilde{f}(\mathbf{X}) = \left(1 + \exp\left(-\sum_{i,j=1}^{d_1,d_2} \mathcal{A}_{ij}\right)\right)^{-1}. \tag{52}$$

We leverage this property to apply context regularization in the binary setting:

Definition B.2 (Context-Regularized Binary Cross-Entropy Loss)**.** Given a training dataset $\mathcal{D}$ and convolutional network f, CR-BCE is defined as:

$$\mathcal{L}_{\text{CRBCE}}(\mathcal{D}) = \mathbb{E}_{(\mathbf{X},y,H)\sim\mathcal{D}}[\ell(\mathbf{X}, y, H)], \quad \text{with}$$

$$\ell(\mathbf{X}, y, H) = \log\left(1 + \exp\left(\sum_{i,j=1}^{d_1,d_2} (-1)^y \mathcal{A}_{ij} H_{ij} + \underbrace{(1 - H_{ij})|\mathcal{A}_{ij}|}_{\text{context regularization}}\right)\right). \tag{53}$$

This definition is motivated by noting that the per-sample binary cross-entropy loss ℓ_{BCE} can be equivalently expressed as:

$$\ell_{BCE}(\mathbf{X}, y) = -y \log(\tilde{f}) - (1 - y) \log(1 - \tilde{f}) = \log(1 + \exp((-1)^y f)), \tag{54}$$

when sigmoid activation is used to obtain class probabilities from logits.

Table 3: **Spawrious Context-Robustness Benchmark.** CR-CE outperforms the standard cross-entropy training, as well as GALS and ABN, across all difficulty levels and context shift settings of the Spawrious dataset.

Spurious Shift Type	CAM Size	Approach	Task Difficulty			Average (↑)
			Easy (↑)	Medium (↑)	Hard (↑)	
One-to-One	7×7	CE (Baseline)	$93.30_{\pm 0.03}$	$83.17_{\pm 0.47}$	$86.58_{\pm 1.88}$	87.69
		GALS	$93.42_{\pm 0.18}$	$84.39_{\pm 2.13}$	$87.11_{\pm 0.27}$	88.31
		EBL	$93.31_{\pm 0.69}$	$83.91_{\pm 1.46}$	$86.44_{\pm 1.45}$	87.89
		CFA	$90.86_{\pm 0.54}$	$79.46_{\pm 0.48}$	$81.77_{\pm 1.73}$	84.03
		Context Masking	$92.36_{\pm 0.70}$	$80.97_{\pm 1.59}$	$83.67_{\pm 3.71}$	85.66
		CR-CE (Ours)	$\mathbf{95.47}_{\pm 0.24}$	$\mathbf{90.51}_{\pm 0.16}$	$\mathbf{91.35}_{\pm 0.34}$	**92.44**
	14×14	CE (Baseline)	$92.83_{\pm 0.01}$	$80.36_{\pm 1.46}$	$87.74_{\pm 0.37}$	86.98
		ABN	$88.01_{\pm 1.49}$	$80.47_{\pm 0.61}$	$80.04_{\pm 2.13}$	82.84
		GALS	$94.72_{\pm 1.32}$	$85.74_{\pm 3.96}$	$90.57_{\pm 0.65}$	90.35
		EBL	$94.20_{\pm 0.28}$	$82.20_{\pm 3.33}$	$87.59_{\pm 1.54}$	88.00
		CFA	$88.98_{\pm 0.14}$	$73.38_{\pm 0.87}$	$75.88_{\pm 1.12}$	79.41
		Context Masking	$93.07_{\pm 0.40}$	$81.79_{\pm 2.29}$	$83.91_{\pm 1.95}$	86.26
		CR-CE (Ours)	$\mathbf{96.35}_{\pm 0.15}$	$\mathbf{92.24}_{\pm 0.30}$	$\mathbf{92.89}_{\pm 0.20}$	**93.83**
Many-to-Many	7×7	CE (Baseline)	$90.86_{\pm 0.54}$	$79.46_{\pm 0.48}$	$81.52_{\pm 1.03}$	83.95
		GALS	$93.47_{\pm 0.36}$	$84.58_{\pm 0.88}$	$78.74_{\pm 0.97}$	85.60
		EBL	$92.43_{\pm 1.31}$	$87.23_{\pm 0.84}$	$81.21_{\pm 0.74}$	86.96
		CFA	$89.14_{\pm 0.89}$	$85.68_{\pm 0.68}$	$81.77_{\pm 1.73}$	85.65
		Context Masking	$92.08_{\pm 1.04}$	$85.03_{\pm 2.37}$	$78.46_{\pm 1.12}$	85.19
		CR-CE (Ours)	$\mathbf{95.55}_{\pm 0.16}$	$\mathbf{91.25}_{\pm 0.18}$	$\mathbf{87.49}_{\pm 0.14}$	**91.43**
	14×14	CE (Baseline)	$88.98_{\pm 0.14}$	$73.38_{\pm 0.87}$	$79.76_{\pm 1.01}$	80.71
		ABN	$89.65_{\pm 0.44}$	$72.81_{\pm 1.32}$	$63.96_{\pm 6.54}$	75.47
		GALS	$95.03_{\pm 1.35}$	$87.63_{\pm 3.80}$	$84.08_{\pm 1.87}$	88.91
		EBL	$93.46_{\pm 0.73}$	$89.05_{\pm 0.47}$	$79.36_{\pm 2.83}$	87.29
		CFA	$87.22_{\pm 0.78}$	$85.68_{\pm 0.68}$	$75.88_{\pm 1.12}$	82.93
		Context Masking	$90.96_{\pm 0.24}$	$85.70_{\pm 2.99}$	$80.80_{\pm 0.71}$	85.82
		CR-CE (Ours)	$\mathbf{96.47}_{\pm 0.06}$	$\mathbf{92.41}_{\pm 0.29}$	$\mathbf{88.66}_{\pm 0.52}$	**92.51**

Table 4: **Hard-ImageNet Benchmark.** Predictions on models trained with CR-CE are least likely to persist when the object is absent from the image. Object-focus increases proportional to the CAM resolution, with some performance cost incurred when the target is present alongside the context.

CAM Size	Approach	No Shift Accuracy (↑)	Accuracy under Target Removal			Rel. Foreground Sensitivity	
			Gray Mask (↓)	Gray BBOX (↓)	Tile (↓)	$\rho = 0.25$ (↑)	$\rho = 0.5$ (↑)
7×7	CE (Baseline)	$95.82_{\pm 0.08}$	$82.4_{\pm 0.13}$	$79.38_{\pm 0.2}$	$77.11_{\pm 0.08}$	$-0.21_{\pm 0.01}$	$-0.36_{\pm 0.01}$
	GALS	$95.91_{\pm 0.15}$	$80.93_{\pm 0.27}$	$76.62_{\pm 1.21}$	$75.69_{\pm 0.73}$	$-0.18_{\pm 0.01}$	$-0.31_{\pm 0.03}$
	EBL	$95.87_{\pm 0.23}$	$77.64_{\pm 1.66}$	$73.20_{\pm 2.07}$	$70.74_{\pm 1.67}$	$-0.11_{\pm 0.04}$	$-0.19_{\pm 0.06}$
	CFA	$\underline{96.00}_{\pm 0.13}$	$82.13_{\pm 0.35}$	$79.24_{\pm 0.50}$	$76.67_{\pm 0.35}$	$-0.20_{\pm 0.01}$	$-0.35_{\pm 0.01}$
	CR-CE (Ours)	$91.42_{\pm 0.28}$	$\mathbf{50.09}_{\pm 0.89}$	$\mathbf{45.78}_{\pm 0.41}$	$\mathbf{46.98}_{\pm 0.2}$	$\mathbf{+0.10}_{\pm 0.02}$	$\mathbf{+0.16}_{\pm 0.01}$
14×14	CE (Baseline)	$95.82_{\pm 0.08}$	$81.2_{\pm 0.61}$	$78.36_{\pm 0.28}$	$77.64_{\pm 0.2}$	$-0.22_{\pm 0.01}$	$-0.34_{\pm 0.003}$
	ABN	$95.24_{\pm 0.54}$	$80.84_{\pm 0.15}$	$73.47_{\pm 1.62}$	$73.33_{\pm 0.23}$	$-0.34_{\pm 0.02}$	$-0.44_{\pm 0.01}$
	GALS	$95.82_{\pm 0.15}$	$78.00_{\pm 3.54}$	$74.27_{\pm 4.35}$	$73.82_{\pm 3.78}$	$-0.12_{\pm 0.08}$	$-0.23_{\pm 0.14}$
	EBL	$95.24_{\pm 0.15}$	$70.31_{\pm 4.46}$	$67.16_{\pm 3.45}$	$65.91_{\pm 5.33}$	$-0.04_{\pm 0.03}$	$-0.05_{\pm 0.14}$
	CFA	$\underline{96.00}_{\pm 0.13}$	$81.42_{\pm 0.50}$	$78.22_{\pm 0.56}$	$76.76_{\pm 0.73}$	$-0.18_{\pm 0.02}$	$-0.36_{\pm 0.06}$
	CR-CE (Ours)	$88.53_{\pm 0.00}$	$\mathbf{37.07}_{\pm 0.48}$	$\mathbf{32.44}_{\pm 0.15}$	$\mathbf{32.76}_{\pm 0.78}$	$\mathbf{+0.17}_{\pm 0.01}$	$\mathbf{+0.23}_{\pm 0.01}$

C Additional Experimental Results

C.1 Spawrious Benchmark on Various Context-Shift Configurations

The Spawrious dataset comprises of four dog breeds spread across six contexts, with three difficulty levels based on the object-context combinations available during training, and two different configurations of context shift: (i) one-to-one shifts, where each class is correlated with one context, and (ii) many-to-many shifts, where groups of classes are correlated with groups of contexts, the latter being observed broadly in practice [23]. We include results on both one-to-one and many-to-many shifts in Table 3. Consistent with our findings in Section 5.3, we observe that CR-CE trained models

Table 5: **Context-Contribution Ratios.** We report the context contribution ratios across each of the assessed approaches. CR-CE trained models place a greater emphasis on object-specific regions to determine predictions and learn to ignore cues from context regions through the process of training.

CAM Size	Approach	Context Contribution Ratio $R\%$ ($\downarrow$)			
		Waterbirds-100	Spawrious[‡]	ImageNet-BG	Hard-ImageNet
7×7	CE (Baseline)	$46.08_{\pm 0.43}$	$44.79_{\pm 0.3}$	$31.83^{\dagger}$	$68.53_{\pm 0.04}$
	GALS	$36.2_{\pm 0.92}$	$28.06_{\pm 1.39}$	$29.75_{\pm 0.5}$	$63.64_{\pm 2.03}$
	EBL	$56.02_{\pm 5.17}$	$21.37_{\pm 1.45}$	$25.35_{\pm 2.06}$	$57.24_{\pm 3.91}$
	CFA	$17.06_{\pm 0.16}$	$43.1_{\pm 2.69}$	$32.49_{\pm 0.07}$	$57.22_{\pm 0.50}$
	Context Masking	$22.48_{\pm 1.72}$	$48.11_{\pm 2.22}$	$32.86_{\pm 0.02}$	$70.63_{\pm 0.06}$
	CR-CE (Ours)	$\mathbf{12.86}_{\pm 0.09}$	$\mathbf{14.89}_{\pm 0.2}$	$\mathbf{14.85}_{\pm 0.00}$	$\mathbf{35.85}_{\pm 0.26}$
14×14	CE (Baseline)	$58.75_{\pm 0.12}$	$47.32_{\pm 0.09}$	$28.27_{\pm 0.27}$	$66.5_{\pm 0.07}$
	ABN	$34.92_{\pm 19.38}$	$49.89_{\pm 16.99}$	$33.37_{\pm 1.01}$	$69.06_{\pm 4.02}$
	GALS	$35.7_{\pm 2.42}$	$17.77_{\pm 2.07}$	$24.14_{\pm 0.13}$	$56.87_{\pm 9.37}$
	EBL	$48.21_{\pm 2.44}$	$29.82_{\pm 1.06}$	$22.94_{\pm 0.83}$	$48.63_{\pm 9.68}$
	CFA	$16.08_{\pm 0.88}$	$46.66_{\pm 0.58}$	$28.42_{\pm 0.02}$	$47.89_{\pm 1.56}$
	Context Masking	$25.24_{\pm 1.27}$	$38.44_{\pm 1.91}$	$27.64_{\pm 0.03}$	$69.35_{\pm 0.2}$
	CR-CE (Ours)	$\mathbf{9.54}_{\pm 0.28}$	$\mathbf{10.98}_{\pm 0.12}$	$\mathbf{9.32}_{\pm 0.06}$	$\mathbf{31.12}_{\pm 0.1}$

[†] No standard deviation is reported; we use pre-trained weights from `torchvision` without modification.
[‡] We report metrics on the hardest subtask and difficulty (M2M-hard) within the Spawrious benchmark.

outperform the baseline and existing explanation-guided approaches in context-shifted accuracy across all difficulty levels.

C.2 Evaluating Context Dependence using Hard-ImageNet

The Hard-ImageNet benchmark [25] assesses model reliance on context by comparing performance on images with and without the target object. Ideally, the model should predict the true class in the presence of the object, and a random class in the absence of the object. In other words, the prediction should not be based on the context alone. The Hard-ImageNet suite also reports Relative Foreground Sensitivity (RFS) [24], which compares accuracy reduction when gaussian noise (with magnitude denoted ρ) is added to either the object or the context region; higher RFS signifies greater object-dependence.

Evaluation of CR-CE on the Hard-ImageNet and comparison with other methods are reported in Table 4. We observe that the CR-CE framework meaningfully outperforms each of the compared methods on key robustness metrics assessed by the Hard-ImageNet benchmark. Example images of Hard-ImageNet demonstrating the effectiveness of CR-CE on suppressing context dependence can be find in Figure 3 and Figure 5, see the classes *Baseball Player*, *Snorkel*, and *Howler Monkey*.

Remark C.1. Hard-ImageNet includes a lot of context-heavy images where object regions comprise only 13.96% of the image on average, leading to a bigger reduction of accuracy in the accuracy-robustness tradeoff.

C.3 Comparison with Explanation-Guided Learning (EGL) Methods

Typically, explanation-guided learning frameworks [14] construct a multi-task optimization objective with a per-sample loss of the following form:

$$\ell_{\mathrm{EGL}} = \ell_{\mathrm{CE}} + \lambda \cdot \ell_{\mathrm{EG}}, \quad \text{with gradient} \quad \nabla \ell_{\mathrm{EGL}} = \nabla \ell_{\mathrm{CE}} + \lambda \nabla \ell_{\mathrm{EG}}. \tag{55}$$

Here, ℓ_{EG} is a secondary explanation-guided goal designed to suppress context contributions. The specific form of this term varies between approaches [38, 34, 20, 29, 11].

For a given (i, j)-patch corresponding to context, gradients of the form in Eq. (55) create a tug-of-war dynamic: the first term seeks to minimize ℓ_{CE} at the cost of ℓ_{EG}, and vice versa. To manage this, the additional hyperparameter λ is introduced and must be carefully tuned to prevent poor classification accuracy. In contrast, CR-CE replaces the mechanism in cross-entropy that increases contributions

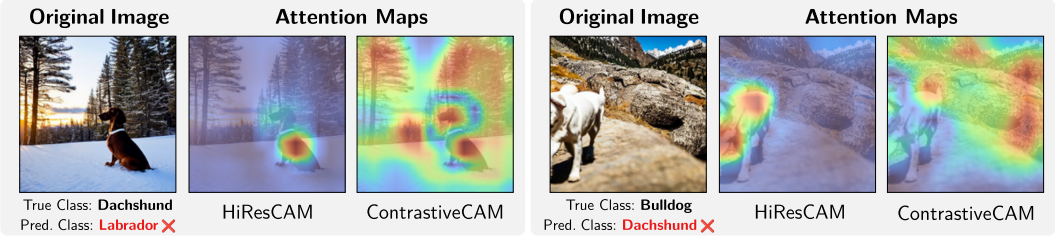


Figure 4: **Superficially Aligned Attribution Maps.** On a ResNet-50 trained with the GALS framework, we observe that HiResCAMs suggest a focus on the object regions; however, the model predicts the context-correlated class observed during training. ContrastiveCAMs faithfully reveal that the context contribution is in fact the primary driver of the prediction. This highlights the essential role that explanation faithfulness plays in ensuring improvements to context robustness in the explanation-guided learning paradigm.

from context regions, preventing competing gradients (taken w.r.t. the specific patch), hence resulting in a more pronounced object-focus effect.

The advantage of this is evident in the Waterbirds-100 dataset. Since it is a binary classification task, the GradCAMs used by GALS coincide with ContrastiveCAMs used by CR-CE. The only difference between the two approaches lies in how the feature attribution method is incorporated in the model training procedure. In this direct comparison, CR-CE continues to outperform GALS, see Table 2.

Explanation Correctness is Key to Robustness Improvements. In Table 2, we observe that some prior explanation-guided methods may even have an inferior performance than the CE training baselines, especially when the tasks become complicated. To investigate this performance drift, we measure the context-contribution ratios for those approaches in Table 5. We observe that models trained by those explanation-guided methods often have greater context dependence, indicating the ineffectiveness of these methods in guiding the feature attributions towards object regions. We hypothesize that this behavior is largely due to their reliance on classical feature attribution method (e.g. Grad-CAM and HiResCAM), which suffer from the unfaithfulness issue theoretically described in Theorem 3.2. Figure 4 shows typical examples where prior explanation-guided methods may be successful in guiding those classical CAMs to focus on the object regions but fail in doing so for the more faithful ContrastiveCAMs.

C.4 CR-CE is Tolerant to Imperfect Object-Context Segmentation

In our experiments with Spawrious and ImageNet datasets in Table 2, we already used AI-predicted object-context segmentation, obtained from SAM3 [5] and MAEs [18], which are approximations and are inevitably imperfect. Here, we further assess CR-CE's tolerance to noise by adding additional noise to these object-context masks. Specifically, we perform random affine transformations (up to 5° of rotation, $\pm 7.5\%$ translation, and $\pm 30\%$ scaling) on each of those masks H. This augmentation results in $\approx 50\%$ mIoU of the original object-context masks to their corrupted counterparts. We train CR-CE using these corrupted masks, and report our results in Table 6. We can see that, even with these noisy object-context segmentation, CR-CE can still predict well with minor degradation in the performance.

C.5 CR-CE is Effective Across Architectures and Scales

Here we provide additional experimental validations of the CR-CE training framework on different network architectures, including EfficientNetV2 [42] and ResNet-18. The results are reported in Table 7 and 8 respectively, demonstrate the effectiveness of CR-CE across variations in architectures and scales.

Table 6: **Context-Shift Robustness using Noisy Masks (NM).** Despite the additional noise introduced to object-context masks, CR-CE remains effective at improving context-shift robustness. CR-CE retains its utility in scenarios where accurate masks are challenging or expensive to obtain.

CAM Size	Approach	Waterbirds-100		ImageNet/ImageNet-BG		Spawrious		
		No Shift (↑)	Shifted (↑)	No Shift (↑)	Shifted (↑)	Easy (↑)	Medium (↑)	Hard (↑)
7×7	CE (Baseline)	$\underline{98.66}_{\pm 0.10}$	$41.74_{\pm 4.15}$	$\underline{82.92}^{\dagger}$	$74.83^{\dagger}$	$90.86_{\pm 0.54}$	$79.46_{\pm 0.48}$	$81.52_{\pm 1.03}$
	CR-CE (Ours) + NM	$92.10_{\pm 0.22}$	$72.61_{\pm 1.26}$	$81.14_{\pm 0.02}$	$74.89_{\pm 0.07}$	$93.07_{\pm 0.32}$	$87.75_{\pm 0.42}$	$82.19_{\pm 0.16}$
	CR-CE (Ours)	$94.38_{\pm 1.21}$	$\mathbf{73.30}_{\pm 4.43}$	$81.85_{\pm 0.03}$	$\mathbf{75.75}_{\pm 0.08}$	$\mathbf{95.55}_{\pm 0.16}$	$\mathbf{91.25}_{\pm 0.18}$	$\mathbf{87.49}_{\pm 0.14}$
14×14	CE (Baseline)	$\underline{98.48}_{\pm 0.09}$	$38.40_{\pm 1.79}$	$82.33_{\pm 0.10}$	$74.72_{\pm 0.06}$	$88.98_{\pm 0.14}$	$73.38_{\pm 0.87}$	$79.76_{\pm 1.01}$
	CR-CE (Ours) + NM	$90.67_{\pm 2.42}$	$76.12_{\pm 3.24}$	$79.83_{\pm 0.06}$	$75.24_{\pm 0.02}$	$93.07_{\pm 0.51}$	$85.64_{\pm 0.17}$	$80.59_{\pm 0.84}$
	CR-CE (Ours)	$94.36_{\pm 0.12}$	$\mathbf{79.13}_{\pm 0.77}$	$81.25_{\pm 0.03}$	$\mathbf{76.84}_{\pm 0.12}$	$\mathbf{96.47}_{\pm 0.06}$	$\mathbf{92.41}_{\pm 0.29}$	$\mathbf{88.66}_{\pm 0.52}$

† No std. dev. is reported; we use pre-trained weights from `torchvision` without modification.

Table 7: **Context-Shift Robustness On EfficientNetV2.** CR-CE improves context-shift robustness on varied convolutional backbones. Following EfficientNetV2's [42] training recipe, we adopt a larger batch size of 4096 and increase the learning rate to 10^{-2} for ImageNet training.

CAM Size	Approach	Waterbirds-100		ImageNet/ImageNet-BG		Spawrious		
		No Shift (↑)	Shifted (↑)	No Shift (↑)	Shifted (↑)	Easy (↑)	Medium (↑)	Hard (↑)
7×7	CE (Baseline)	$91.08_{\pm 0.05}$	$39.31_{\pm 3.01}$	$\underline{83.71}^{\dagger}$	$73.58^{\dagger}$	$92.62_{\pm 0.41}$	$81.49_{\pm 0.29}$	$77.50_{\pm 0.91}$
	CR-CE (Ours)	$\underline{94.31}_{\pm 1.93}$	$\mathbf{78.55}_{\pm 5.01}$	$83.36_{\pm 0.16}$	$\mathbf{74.46}_{\pm 0.08}$	$\mathbf{93.86}_{\pm 0.62}$	$\mathbf{84.23}_{\pm 0.67}$	$\mathbf{88.29}_{\pm 0.22}$
14×14	CE (Baseline)	$84.64_{\pm 0.14}$	$38.88_{\pm 1.31}$	$\underline{84.00}_{\pm 0.07}$	$74.53_{\pm 0.07}$	$93.11_{\pm 0.42}$	$80.27_{\pm 2.06}$	$79.30_{\pm 1.17}$
	CR-CE (Ours)	$\underline{94.02}_{\pm 1.15}$	$\mathbf{80.43}_{\pm 5.01}$	$83.67_{\pm 0.03}$	$\mathbf{75.20}_{\pm 0.04}$	$\mathbf{94.84}_{\pm 0.33}$	$\mathbf{86.98}_{\pm 0.04}$	$\mathbf{88.85}_{\pm 0.55}$

† No std. dev. is reported; we use pre-trained weights from `torchvision` without modification.

D About Data Augmentation with Object-Context Mixing

A straightforward yet often effective approach to improving robustness to context shifts is to augment the training dataset with synthetic images that combine objects with diverse contexts [12]. This approach is orthogonal to the methods considered in this paper, including explanation-guided methods, as it leaves the training procedure unchanged and simply introduces additional augmented training samples. Therefore, it can be readily combined with our proposed method. However, its effectiveness may be limited by the finite set of object–context combinations observed during training, potentially leaving the model vulnerable to unseen context shifts at evaluation time, particularly adversarial ones. In contrast, our proposed training framework suppresses the influence of context regions, not limited by the artificial object–context combinations seen during training, potentially providing greater robustness to out-of-distribution scenarios.

Nevertheless, we empirically evaluate this approach by augmenting each training image with an additional image in which the object is paired with a context uniformly sampled from the non-target classes, thereby doubling the size of the training dataset. We report the effect of pairing this augmentation with both CE and CR-CE in Table 9. We can see that the data augmentation technique performs well in simple tasks, such as Waterbirds, but becomes worse when the task becomes harder.

E Training Details

Hyperparameter Settings. For all methods, we use Momentum SGD [32] with a momentum parameter 0.9. The hyperparameter settings shown in Table 10 are applied across all methods. GALS [29], ABN [11] and Energy loss [34] each have an additional hyperparameter λ which weights the regularization strength of spatial feature attributions, which we keep at their default value of 1.

Computational Resources. We train each of our models using a single NVIDIA H100 GPU. The most efficient loss function, cross-entropy, took a maximum of 5.5 hours to train ResNet-50 on ImageNet, and a minimum of 20 minutes on the Waterbirds dataset. Training with CR-CE took a maximum of 6.1 hours per ResNet-50 run on the ImageNet dataset, and a minimum of 20 minutes on the Waterbirds dataset. Each of the other methods approximately took between 8 and 12 hours to train a ResNet-50 on ImageNet. Each of the methods used a maximum of 65 GB of VRAM during training on the ImageNet dataset, with gradient accumulation applied as necessary to stay within the VRAM limit of the GPU. We trained a total of 582 models: 3 independent runs for each approach

Table 8: **Context-Shift Robustness On ResNet-18.** CR-CE improves context-shift robustness on the smaller ResNet-18 backbone. We observe a more pronounced trade-off between in-distribution and context-shifted accuracy due to the smaller size of the backbone. In contrast, this trade-off is minimized by the larger capacity of the EfficientNetV2 backbone, see Table 7.

CAM Size	Approach	Waterbirds-100		ImageNet/ImageNet-BG		Spawrious		
		No Shift (↑)	Shifted (↑)	No Shift (↑)	Shifted (↑)	Easy (↑)	Medium (↑)	Hard (↑)
7×7	CE (Baseline)	$79.96_{\pm 12.03}$	$38.00_{\pm 3.79}$	$\underline{73.49}^{\dagger}$	$58.38^{\dagger}$	$80.89_{\pm 0.92}$	$59.34_{\pm 0.73}$	$52.16_{\pm 1.27}$
	CR-CE (Ours)	$\underline{88.48}_{\pm 0.09}$	$\mathbf{53.55}_{\pm 10.44}$	$72.40_{\pm 0.05}$	$\mathbf{62.61}_{\pm 0.02}$	$\mathbf{90.77}_{\pm 0.07}$	$\mathbf{75.93}_{\pm 0.56}$	$\mathbf{68.99}_{\pm 0.29}$
14×14	CE (Baseline)	$\underline{97.59}_{\pm 0.36}$	$30.41_{\pm 8.90}$	$\underline{75.02}_{\pm 0.10}$	$60.43_{\pm 0.06}$	$74.84_{\pm 2.08}$	$46.14_{\pm 2.45}$	$37.81_{\pm 1.14}$
	CR-CE (Ours)	$88.25_{\pm 0.22}$	$\mathbf{59.95}_{\pm 7.97}$	$71.44_{\pm 0.06}$	$\mathbf{63.37}_{\pm 0.04}$	$\mathbf{91.64}_{\pm 0.26}$	$\mathbf{79.42}_{\pm 0.23}$	$\mathbf{70.68}_{\pm 1.44}$

† No std. dev. is reported; we use pre-trained weights from `torchvision` without modification.

Table 9: **Context-Shift Robustness using Object-Context Augmentation.** Generating object-context combinations using copy-paste augmentation only improves robustness in the Waterbirds setting, when object-context augmentation simulates the test set. In larger and more realistic context-shift scenarios, augmenting contexts does not produce enough signal to suppress context contributions.

CAM Size	Approach	Waterbirds-100		ImageNet/ImageNet-BG		Spawrious		
		No Shift (↑)	Shifted (↑)	No Shift (↑)	Shifted (↑)	Easy (↑)	Medium (↑)	Hard (↑)
7×7	CE (Baseline)	$\underline{98.66}_{\pm 0.10}$	$41.74_{\pm 4.15}$	$\underline{82.92}^{\dagger}$	$74.83^{\dagger}$	$90.86_{\pm 0.54}$	$79.46_{\pm 0.48}$	$81.52_{\pm 1.03}$
	CE + OCM	$96.01_{\pm 0.11}$	$90.64_{\pm 0.81}$	$80.38_{\pm 0.13}$	$72.80_{\pm 0.04}$	$79.65_{\pm 0.39}$	$49.15_{\pm 2.01}$	$42.99_{\pm 1.01}$
	CR-CE (Ours)	$94.38_{\pm 1.21}$	$73.30_{\pm 4.43}$	$81.85_{\pm 0.03}$	$\mathbf{75.75}_{\pm 0.08}$	$\mathbf{95.55}_{\pm 0.16}$	$\mathbf{91.25}_{\pm 0.18}$	$\mathbf{87.49}_{\pm 0.14}$
	CR-CE (Ours) + OCM	$95.35_{\pm 1.02}$	$\mathbf{92.63}_{\pm 0.97}$	$79.52_{\pm 0.03}$	$73.79_{\pm 0.01}$	$94.40_{\pm 0.09}$	$87.95_{\pm 0.61}$	$76.96_{\pm 0.60}$
14×14	CE (Baseline)	$\underline{98.48}_{\pm 0.09}$	$38.40_{\pm 1.79}$	$82.33_{\pm 0.10}$	$74.72_{\pm 0.06}$	$88.98_{\pm 0.14}$	$73.38_{\pm 0.87}$	$79.76_{\pm 1.01}$
	CE + OCM	$95.51_{\pm 0.22}$	$92.73_{\pm 1.06}$	$81.08_{\pm 0.14}$	$73.70_{\pm 0.05}$	$70.69_{\pm 2.47}$	$40.84_{\pm 0.74}$	$32.17_{\pm 5.06}$
	CR-CE (Ours)	$94.36_{\pm 0.12}$	$79.13_{\pm 0.77}$	$81.25_{\pm 0.03}$	$\mathbf{76.84}_{\pm 0.12}$	$\mathbf{96.47}_{\pm 0.06}$	$\mathbf{92.41}_{\pm 0.29}$	$\mathbf{88.66}_{\pm 0.52}$
	CR-CE (Ours) + OCM	$94.88_{\pm 0.29}$	$\mathbf{93.36}_{\pm 0.88}$	$78.85_{\pm 0.07}$	$74.75_{\pm 0.03}$	$95.30_{\pm 0.24}$	$88.55_{\pm 0.36}$	$77.51_{\pm 0.41}$

† No std. dev. is reported; we use pre-trained weights from `torchvision` without modification.

Table 10: **Hyperparameters** are dataset-specific and kept consistent across approaches. Hard-ImageNet and ImageNet-1k are trained for fewer epochs since we fine-tune ImageNet-trained models. We increase the batch size as number of classes (2, 4, 15, and 1000 respectively) increases.

Dataset	Epochs	Batch Size	LR	Momentum	Weight Decay	Scheduler
Waterbirds-100	150	64	10^{-3}	0.9	10^{-4}	-
Spawrious	150	128	10^{-3}		10^{-4}	-
Hard-ImageNet	100	256	$5 \cdot 10^{-3}$		10^{-4}	-
ImageNet-1k / ImageNet-BG	20	512	$5 \cdot 10^{-4}$		10^{-5}	Cosine Annealing

and backbone on 9 different training sets (6 for Spawrious, 1 each for Waterbirds, ImageNet-BG, and Hard-ImageNet), with 2 CAM resolutions for all but ABNs. ABN increases the CAM resolution as part of its architecture design; we do not reduce it to 7×7.

Reproducibility. The source code for this project can be accessed at `https://codeberg.org/jinensetpal/context-regularized-cross-entropy/`.

Asset Licenses. We used the following existing assets, listed with their respective licenses:

- The Waterbirds dataset [37] is licensed under MIT.
- The Spawrious dataset [23] is licensed under CC0-1.0.
- The ImageNet dataset [8] is licensed only for non-commercial purposes.
- The ImageNet-BG dataset [30] is published under a public domain license.
- The Hard-ImageNet dataset [25] is published under CC-0.
- Model weights for the Masked AutoEncoder (MAE) [18] trained on the Large-Scale Unsupervised Segmentation (LUSS) [13] dataset is licensed under CC BY-NC 4.0.
- Model weights for ABNs [11] is licensed under MIT.
- SAM3 [5] is published under a custom SAM License which permits the academic use of their tool.
- The `torchvision` library is published under a BSD 3-Clause License.

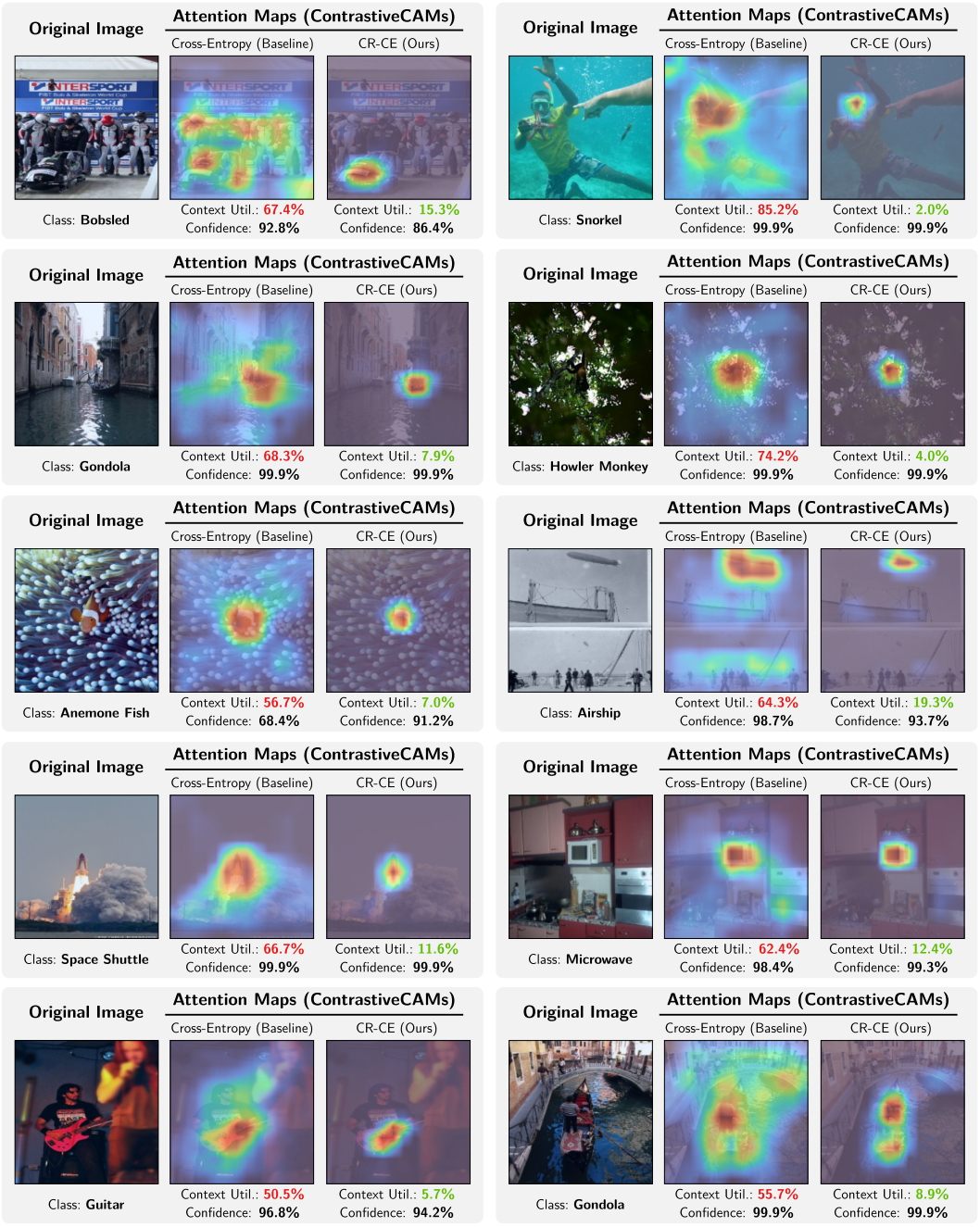


Figure 5: **ImageNet-BG** and **Hard-ImageNet.** Extended qualitative evaluations of object-context dependence by cross-entropy and CR-CE trained models. Images belonging to the *Snorkel* and *Howler Monkey* classes correspond to the Hard-ImageNet validation set, whereas the rest correspond to the ImageNet-BG test set. Across each of the examples, we observe the context dependence in models trained using cross-entropy loss is significantly mitigated by training with CR-CE.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We aim for factual correctness and clarity in both the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6, we discuss the limitations of our approach.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 2 discusses the setting that we consider, and Appendix A includes proofs for all theoretical claims. To the best of our knowledge, each each proof is complete and correct.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Code and data notwithstanding, one may reproduce our results by training models using CR-CE (Definition 4.1) in place of cross-entropy. Our submission includes the necessary details to reproduce our results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We include the code to reproduce our experiments in the supplemental material, and will open-source it under a permissive license in the camera-ready version. The datasets we use are public and freely accessible.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include these details in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For each approach that we assess, we compute metrics from 3 independent training runs and report the mean and standard deviation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and, to the best of our knowledge, believe that our work conforms to the guidelines in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We motivate our approach with societal impacts in Section 1 and restate it briefly in Section 6.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Detailed in Appendix E.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will publish our code alongside corresponding documentation under a permissive license as part of our released assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not use human subjects or crowdsourced experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not use human subjects or crowdsourced experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We use LLMs as a language tool to edit the paper only. LLMs were not used for writing code or influencing the methodology in any way.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.