Consistent Explanations by Contrastive Learning

Vipin Pillai
University of Maryland, Baltimore County
vp7@umbc.edu

Soroush Abbasi Koohpayegani
University of Maryland, Baltimore County
soroush@umbc.edu

Ashley Ouligian
Northrop Grumman
Ashley.Rothballer@ngc.com

Dennis Fong
Northrop Grumman
Dennis.Fong@ngc.com

Hamed Pirsiavash
University of California, Davis
hpirsiav@ucdavis.edu

Abstract

Post-hoc explanation methods, e.g., Grad-CAM, enable humans to inspect the spatial regions responsible for a particular network decision. However, it is shown that such explanations are not always consistent with human priors, such as consistency across image transformations. Given an interpretation algorithm, e.g., Grad-CAM, we introduce a novel training method to train the model to produce more consistent explanations. Since obtaining the ground truth for a desired model interpretation is not a well-defined task, we adopt ideas from contrastive self-supervised learning, and apply them to the interpretations of the model rather than its embeddings. We show that our method, Contrastive Grad-CAM Consistency (CGC), results in Grad-CAM interpretation heatmaps that are more consistent with human annotations while still achieving comparable classification accuracy. Moreover, our method acts as a regularizer and improves the accuracy on limited-data, fine-grained classification settings. In addition, because our method does not rely on annotations, it allows for the incorporation of unlabeled data into training, which enables better generalization of the model. Our code is available here: https://github.com/UCDvision/CGC

1. Introduction

Deep neural networks have become ubiquitous in many applications owing to their performance on several computer vision tasks. Although they have been instrumental in achieving state-of-the-art accuracy, deep neural networks are widely considered to be black box systems, which is not desirable. For example, if an AI system is deployed to identify a malignant tumor from CT scans, it is important for medical experts to understand the reasoning behind the decision-making process [39]. This not only enables building trust, but also helps identify any spurious correlations

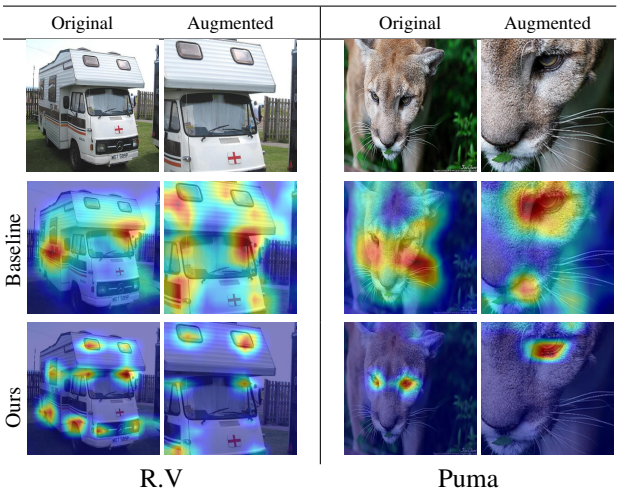

Figure 1. Our method significantly improves the consistency of Grad-CAM explanation heatmaps under data augmentation. For both the RV and the Puma, our method highlights the same portions of the image in both the original and augmented versions.

that the network may have inadvertently learned to use to make its decision [34]. In recent years, there have been attempts to open this black box by designing frameworks to explain the network's decision-making process. Post-hoc explanation methods such as CAM [46], Grad-CAM [32], and Full-Grad [36] generate a heatmap in the size of the image with higher values corresponding to the regions that contributed most to the network's decision.

Unlike image labels, there can be multiple valid explanations for a given image category. Furthermore, a valid explanation might not involve the entire object-segmentation area. For example, in order to correctly classify an image of a dog, the network might rely on just the facial features of the dog, or the texture of the fur and the tail, or a combination of both. Each of these are valid explanations, and hence, generating ground truth annotations for explanations

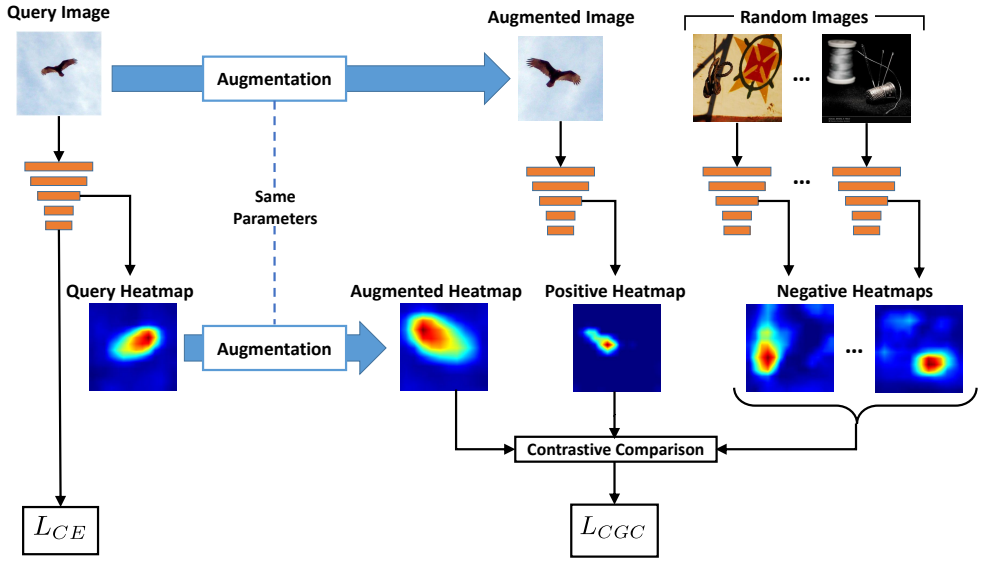

Figure 2. **The block diagram of our method.** Our method consists of both cross-entropy loss (L_{CE}) and contrastive Grad-CAM consistency loss (L_{CGC}). We load a batch of random images, and consider one to be the query image. We feed the query image to the network and calculate L_{CE}. We calculate Grad-CAM for this image on the top predicted category and then augment the heatmap. We then augment all the images in the batch, feed them to our model, and calculate the Grad-CAM heatmap for the top predicted category. The top category is chosen using the original image and not the augmented one. (Note that all random images are also augmented with independently sampled parameters. We do not show this to reduce the clutter.) The heatmap from the augmented query image is considered to be the positive example. The heatmaps from the other random images in the batch are the negative examples. The augmented query heatmap, the positive heatmap, and the negative heatmaps are all fed into our contrastive comparison function to produce our L_{CGC} term.

is a not well defined task. This makes it difficult to directly supervise the network's explanation during training.

It is shown that most interpretation methods are not consistent with spatial transformation of the images. For instance, shifting an image does not shift the interpretation heatmap in the same way [11, 18]. In addition, in fine-grained visual categorization, it is important to learn the subtle yet discriminative features across classes (e.g., wing color, beak and eyes for a bird) [42] and hence the network interpretation should focus on the most salient features that discriminate the correct class from other classes. Assuming Grad-CAM is a truthful interpretation method, we are interested in improving the training process of the deep network so that its Grad-CAM interpretation is more consistent with respect to spatial transformations, thus making the model more interpretable. We use Grad-CAM as the key interpretation method for the rest of the paper since it passes the sanity check introduced in [1] and is end-to-end differentiable.

Inspired by self-supervised learning, we argue that a spatial affine transformation on an image should correspond to such transformation in the interpretation. For instance, given image of a "dog", the heatmap for "dog" category should highlight the dog and it should shift or zoom if we shift or zoom into the image. Figure 1 shows an example where the change in Grad-CAM heatmap for the baseline network computed for the "RV" and "Puma" categories is not consistent with the spatial affine transformation applied to the images. We adopt ideas from recently developed contrastive self-supervised learning literature [12, 14, 24] and design a loss function that encourages the Grad-CAM of an image to be close to the Grad-CAM of an augmented version of the same image while being far from the Grad-CAM of other random images.

We evaluate our CGC method for both classification accuracy and quality of explanations measured by the "Content Heatmap" (CH) metric introduced in [28]. A similar metric was also introduced in [37] to measure the energy of explanation heatmap inside an adversarial patch. CH measures the cumulative heatmap contained within an annotated object mask and is a proxy to measure the consistency of an explanation heatmap with respect to human annotations. Moreover, since our method guides the model to focus mainly on the most discriminative features of the images, it improves the accuracy in fine-grained classification settings while improving the consistency of interpretations. The effect is even more prominent in learning with fewer labels on fine-grained tasks. We believe this is due to the regularization effect that our method adds to the training process. In addition, since our loss function does not need

labeled data, it can benefit from unlabeled data during training. Our experiments show that our method improves the consistency of interpretation with respect to human annotations as well as the classification accuracy in limited labels and fine-grained settings.

2. Related Work

Explanation Methods: Early methods used to explain the decision-making process of a deep neural network have relied on developing post-hoc explanation frameworks that generate explanation heatmaps for a given network, input image, and the corresponding predicted class. [33] introduced a method that measured the gradient of the predicted class with respect to the input image and used spatial locations with large gradient magnitudes to obtain a saliency map. This was improved upon by [35, 38] to obtain sharper saliency maps. Such gradient-based saliency methods have been shown to match the state-of-the-art example-based explanations on several datasets [30]. Class Activation Mapping (CAM) [46] was introduced to generate a coarse localization heatmap by using a global average pooling (GAP) layer to compute the gradients flowing into the final convolution layer. Gradient-weighted Class Activation Mapping (Grad-CAM) [32] generalized CAM by eliminating the need for a GAP layer to compute coarse localization heatmaps. Another class of explanation methods perturb the input to the model and observe the resulting changes to the output [5, 8, 9, 26, 43].

Self-Supervised Learning: Self-supervised learning methods rely on introducing pretext tasks based on structural priors of the data without using image labels to learn meaningful representations. Spatial structure in the visual domain [7, 23, 24], color information [20, 45], and spatial orientation [10] have been used to design pretext tasks. Recently, such hand-crafted pretext tasks have been superseded by methods that instead learn representations by contrasting the feature vectors of positive pairs with those of negative pairs [3, 14]. We build upon these ideas by leveraging contrastive learning on the network's explanation heatmaps instead of the representations. This aligns with our prior that the network's explanation for a given image should be consistent under augmentations of that image.

Consistency for Explanations: [29] use domain knowledge to align explanations with prior knowledge in the form of importance scores. [13] use adversarial perturbation on the input images for explanation consistency with the original image. [40] use causal masking to remove salient regions of the input image and generate positive and negative contrast images to improve model interpretability. [16, 17] propose contrastive learning to improve interpretability for NLP models. [11] introduced the idea of imposing a perceptual consistency prior on the attention heatmaps while training the network for multi-label image classification. The key idea in [11] is that the CAM [46] attention heatmap of an image should follow the same transformation if the image is transformed. A similar idea to enforce consistency regularization on the CAM attention heatmaps using the concept of attention separability and cross-layer attention consistency [41] was introduced for the task of weakly-supervised semantic segmentation.

Our work differs from these in that we use a set of negative examples along with the standard image transformations in a contrastive setting. Negative examples play an important role in ensuring that the interpretation relies on image-specific regions rather than being biased towards a blob in the middle of the image or spread around the image uniformly. Moreover, most of these works evaluate the interpretation by using it as a semantic segmentation tool. However, we believe that an interpretation heatmap should not necessarily highlight the whole object mask. Instead, it should highlight the "most discriminative regions" of an image, which is a strict subset of the semantic segmentation mask. We emphasize that we are not introducing a new explanation method, but rather learning a model that is tuned to be explainable for a given explanation method.

Probably, [28] is the closest to our work as it uses self-supervised learning ideas to remove spurious correlations in the interpretation heatmap. [28] uses a different self-supervised pseudo task for reducing the contextual reasoning by feeding synthetic composite images during training which is out-of-distribution data. Our work is focused on contrastive learning which has driven the recent progress in self-supervised learning community and uses in-distribution images only during training. Unlike [28], our method does not require labels to encourage explanation consistency and hence we are able to leverage additional unlabeled data for our contrastive explanation consistency loss together with the labeled data used for the standard cross-entropy loss.

3. Method

Figure 2 shows a block diagram of our method. Our training process consists of both categorical cross-entropy loss (L_{CE}) and contrastive Grad-CAM consistency loss (L_{CGC}). In this section, we first present a brief overview of the Grad-CAM interpretation algorithm and then describe the contrastive Grad-CAM consistency loss term.

Background on Grad-CAM [32]: Consider an input image x and a deep neural network f. Let y be the vector of output logits when we feed x to the model f where, y^t corresponds to the output for category t. The Grad-CAM of model f for image x and a given category t is a heatmap that highlights the regions of image x responsible for the model's classification of the image as category t. We calculate this heatmap by choosing an intermediate convolutional layer and then linearizing the rest of the network to be interpretable. More specifically, we calculate the derivative

of the predicted output with respect to each channel of the convolutional layer averaged over all spatial locations. This results in a scalar for each channel that captures the importance of that channel in making the current prediction. Then, we calculate a weighted average of all activations of the convolutional layer with the above importance weights for each channel to get a 2D matrix over spatial locations. Finally, we keep only positive numbers and resize it to the size of input image to get the interpretation heatmap.

3.1. Contrastive Grad-CAM Consistency Loss

We are interested in training the image classification model so that its Grad-CAM heatmaps are consistent with spatial transformations. For instance, when we shift the image, the interpretation heatmap should also shift in the same way. Also, the heatmap should be specific to the image, as in not always focused on a blob in the middle of the image or be spread around the whole image.

Inspired by contrastive learning methods in self-supervised learning, we design a contrastive loss function for the interpretation heatmaps that acts as a regularizer when added to the standard cross entropy loss for supervised learning. We want the transformed interpretation of a query image to be close to the interpretation of the transformed query image while being far from interpretations of other random images.

More formally, we assume $g(.)$ is the Grad-CAM operator that calculates the interpretation heatmap for the top predicted category of an input image. Given a set of n random images $\{x_j\}_{j=1}^n$, we augment them with independent random spatial transformations $T_j(.)$ which involves a combination of random scaling, cropping, and flipping. This is similar to the standard augmentation usually done in deep learning. Then, we feed the augmented images through the model and calculate their Grad-CAM heatmaps to get $\{g_j(T_j(x_j)\}_{j=1}^n$. We assume one of the images x_i where $i \in 1..n$ is the query image and calculate its Grad-CAM heatmap without any transformation. We then apply the same transformation we had applied to x_i to the Grad-CAM heatmap, instead of the image, to get $T_i(g_i(x_i))$.

Our main idea is that if we transform an image, the interpretation should also be transformed in the same way. In addition, the interpretation should be specific to each image. We want $T_i(g_i(x_i))$ to be close to $g_i(T_i(x_i)))$ and far from $\{g_j(T_j(x_j)\}_{j\neq i}$. Hence, we define the following loss function:

$$L_i = -\log \frac{\exp\big(\text{sim}\big(T_i(g_i(x_i)), g_i(T_i(x_i))\big)/\tau\big)}{\sum_{j=1}^n \exp\big(\text{sim}\big(T_i(g_i(x_i)), g_j(T_j(x_j))\big)/\tau\big)}$$

where τ is the temperature hyperparameter and $\text{sim}(a, b)$ measures a similarity between two heatmaps. In our experiments, we use cosine similarity. Note that cosine similarity is equivalent to L2 distance metric on normalized features. $g_i(.)$ always calculates the Grad-CAM of the top prediction of the original image regardless of the transformation since it is important to keep the category that the Grad-CAM heatmap is calculated on consistent for the positive pair involving the query image x_i.

We call our loss term Contrastive Grad-CAM Consistency Loss (L_{CGC}). This loss is similar to the standard contrastive self-supervised learning loss [14] with two main differences: (1) Our loss is defined on the interpretation of the network output instead of the image features; (2) The interpretation of the original query image is also augmented with the same parameters to match the interpretation of the augmented image. This compensates for the transformation to make the interpretations aligned.

In practice, since we run the optimization on mini-batches, we assume each image is the query once and sum over all losses optimizing $L_{CGC} = \sum_i L_i$. This can be implemented efficiently for the whole mini-batch by augmenting each image once and calculating Grad-CAM for each image twice. Some contrastive self-supervised learning algorithms like MoCo [14] use a memory bank to increase the number of negative pairs, but for simplicity, we do not use a memory bank and use mini-batches of size 256. Thus, our method is more similar to SimCLR [4] than MoCo [14].

Our final loss is the combination of the standard cross-entropy loss (L_{CE}) and our contrastive Grad-CAM consistency loss (L_{CGC}). Hence, we minimize the following loss function:

$$L = L_{CE} + \lambda L_{CGC}$$

where, λ is a hyper-parameter that controls the trade-off between the two loss terms. Note that our L_{CGC} loss term does not use image labels as it uses pseudo labels for Grad-CAM. This enables us to use additional unlabeled data to improve both the accuracy and explainability of the resulting model in Section 4.3.

4. Experiments

In this section, we perform a variety of experiments using our CGC method. For each of the experiments, 'baseline' refers to a model trained from scratch using the standard cross-entropy loss unless noted otherwise. We will report the classification accuracy along with the following metrics used for evaluating the explanation heatmaps:

Content Heatmap (CH): Introduced in [28], this metric is a measure of the summation of ℓ_1-normalized heatmap contained within the annotated bounding box of the object. If the model interpretation is consistent with human annotations of the object location, we can assume that the percentage of the heatmap that lies inside the object annotation mask should be close to 100%. Hence, we expect this metric to be high.

CGC loss: We also evaluate the explanation heatmaps using the same L_{CGC} loss that we use for training our models. Although this loss is already used in our optimization, we believe it is important to show that the loss is small on the unseen test data as well, i.e., the method generalizes from the training to test set. We use ImageNet [31] validation set to report this metric. We use a batch size of 32 to compute this loss term. For every query image in the batch, we pair it with an augmentation of the corresponding query image as the positive pair and consider each of the remaining 31 images in the batch to be the negative pairs.

Insertion AUC score (IAUC): This metric [26] successively inserts pixels from the highest to lowest attribution scores and makes predictions. The area under the curve defined by the prediction scores is then defined as the IAUC score. We expect the IAUC score to be higher for a better interpretation.

Implementation Details: We use PyTorch [25] to train and evaluate our models for all experiments. We use SGD (weight decay=$1e-4$, learning rate=0.1, momentum=0.9, and batch size=256) to optimize both ResNet18 and ResNet50 [15] models. We train ImageNet [31] models for 90 epochs and decay the learning rate by 0.1 every 30 epochs. For transformation T_i in our loss term, we use standard data augmentations (scaling, flipping, and translation). The baseline models also use these same augmentations. We use ResNet50 architecture for all our experiments unless otherwise noted. Training our ResNet50 on 2 Titan-RTX GPUs takes approximately 100 hours, whereas training the baseline takes approximately 70 hours on the same setting. We use $\tau = 0.5$ for all experiments using our method. For ImageNet dataset, we use $\lambda = 1.0$ for our method using ResNet18 and $\lambda = 0.5$ for ResNet50.

We first report our results on ImageNet and UnRel [27] datasets. We then show results on fine-grained classification tasks including limited data settings. Finally, we discuss results using unlabeled data with our loss term.

4.1. Contrastive Grad-CAM Consistency (CGC)

We train a network from scratch using the CGC method (cross-entropy loss and contrastive loss) and compare it to a baseline network trained from scratch using cross-entropy loss only. We report the Top-1 classification accuracy as well as heatmap evaluation metrics. We show results for both the ImageNet and UnRel [27] datasets in order to show that our approach generalizes across multiple datasets. For both datasets, on the ResNet50 architecture, we show that our model has a slight (less than 2% points) decrease in classification accuracy but shows a significant (greater than 15% points) improvement on the explanation metrics.

ImageNet: Quantitative results on ImageNet validation set are reported in Table 1. For the ResNet50 model, we have a marginal drop in classification accuracy (1.5 points), whereas CH increases by 17 points. Since, ImageNet is a large dataset, we do not expect our regularizer to improve the classification accuracy on ImageNet. Moreover, our main focus is on improving the consistency of the explanations and we are willing to accept a marginal drop in classification accuracy at the cost of improved explanation consistency.

Table 2 reports the evaluation results using the IAUC metric [26] on the ImageNet validation set and shows that our method quantitatively improves the Grad-CAM explanation heatmaps of the underlying model.

Arch	Method	Top-1 Acc (%)	CH (%)	CGC Loss
ResNet18	Baseline	**69.76**	54.47	3.19
	GCC [28]	67.74	57.73	3.14
	Ours (CGC)	66.37	**65.83**	**2.59**
ResNet50	Baseline	**76.13**	54.77	3.15
	GCC [28]	74.40	59.42	3.09
	Ours (CGC)	74.60	**71.75**	**2.64**

Table 1. Classification accuracy along with the Content Heatmap (CH) and CGC Loss explanation metrics on ImageNet validation set. Note that lower is better for CGC Loss.

Method	Insertion AUC
Baseline	0.4860
Ours (CGC)	**0.5216**

Table 2. The Grad-CAM explanation maps generated by our ResNet50 model outperforms the baseline ResNet50 model on the IAUC metric [26] using the ImageNet validation set.

UnRel: The UnRel dataset [27] consists of images capturing unusual relations between objects. These images were collected from the web using triplet queries such as 'person ride giraffe'. This dataset thus captures objects occurring in unusual spatial configurations and contexts. Both the baseline and CGC models are trained on ImageNet and evaluated on the 28 object categories of the UnRel dataset that overlap with ImageNet. We report the corresponding quantitative results in Table 3. We observe that although our CGC model was trained on ImageNet, which is an object centric dataset, the improvement in our explanation heatmap generalizes to objects occurring in unusual backgrounds and spatial configurations.

Generalization to another interpretation method: We also show that the improved explanations of our model trained using Grad-CAM generalize to Contrastive Top-down Attention (cMWP) [44], which is another explanation method. We report CH evaluation results using cMWP in Table 4.

Model	Top-1 Acc (%)	CH (%)
Baseline	**40.66**	51.66
Ours (CGC)	38.25	**74.20**

Table 3. Classification accuracy and Content Heatmap (CH) evaluation on the 28 UnRel categories [27] that overlap with ImageNet.

Method	CH (%) using cMWP [44]
Baseline	74.78
GCC [28]	75.08
Ours (CGC)	**75.50**

Table 4. Our method generalizes to another explanation method, Contrastive Top-down Attention (cMWP) [44] although our model was trained using Grad-CAM. We report the CH metric computed using cMWP on ImageNet validation set. All models use ResNet50 architecture.

Role of negative heatmaps in L_{CGC}: As part of our L_{CGC} loss, together with an augmented image as the positive pair, we use the rest of the images in the batch to compute the negative heatmaps. The presence of negative heatmaps is expected to encourage the model to learn explanation heatmaps specific to each image. This would result in the model learning features corresponding to the most discriminative regions of the object. If the negative heatmaps are not used as part of the contrastive loss, the model can cheat by learning a trivial solution for the explanation consistency and would result in the explanation heatmaps being spread uniformly across the image. We verify this be training a model where the L_{CGC} uses a simple ℓ_2 loss on the normalized heatmap of the positive pair and does not use negative heatmaps. We observe that the resulting model indeed results in explanation heatmaps which are diffused across the image as verified by the low CH in Table 5.

Model	ResNet18		ResNet50	
	Top-1 Acc (%)	CH (%)	Top-1 Acc (%)	CH (%)
CGC	66.37	**65.83**	74.60	**71.75**
CGC w/o neg	67.00	44.08	74.80	39.84

Table 5. Comparison of ImageNet evaluation against model trained without using negative heatmaps as part of L_{CGC}. The model trained without the negative heatmaps as part of L_{CGC} loss results in a very low CH, thus confirming our hypothesis that the lack of negative heatmaps would result in a model learning a trivial solution of generating heatmaps diffused throughout the image.

4.2. Fine-Grained Classification

We now report the results for fine-grained classification on CUB-200 [42], FGVC-Aircraft [21], Stanford Cars-196 [19], and VGG Flowers [22] datasets. All models used for fine-grained classification are pretrained on ImageNet and all layers are fine-tuned for the new dataset.

We find that our method can be used as a form of regularization, which is particularly helpful in these fine-grained classification scenarios. While our method trains to produce better interpretation, using a contrastive loss on the interpretation will allow the model to learn unique attention for individual classes. Note that the negative samples in the contrastive loss encourages the interpretation to be different from other samples. As a result, the trained model limits the attention to the most discriminative part(s) of the object and hence, our method acts like a regularizer when adopted for fine-grained classification tasks.

For our CGC method, we use $\lambda = 0.25$ for the CUB-200 and Cars-196 datasets, and $\lambda = 0.5$ for the FGVC-Aircraft and VGG Flowers datasets.

Results for classification accuracy are shown in Table 6, and results for Content Heatmap evaluation are shown in Table 8. For all fine-grained datasets, the model trained with our CGC method achieves both improved classification accuracy and improved explanation scores. For most datasets, improvements to classification accuracy is marginal, but the FGVC-Aircraft dataset shows an improvement of over 2% points. This experiment demonstrates that our method not only improves explainability, but also classification accuracy for datasets that require the network to focus on the "most discriminative" features of an object.

Limited Training Data: Using our method for regularization can be particularly helpful when the amount of training data is limited. We evaluate our method with the same fine-grained classification settings, but limit the amount of training data for each class. We evaluate 1, 5, and 10-shot training settings. We repeat our experiment for 20 episodes and report mean and standard deviation over all episodes. In each episode, we randomly select $n \in \{1, 5, 10\}$ samples from each class as training set and use the rest of the samples as a validation set. We initialize the model from a ResNet50 model pretrained on ImageNet with categorical cross-entropy loss only, then finetune all layers on the limited training set. We use the same random seed for our method as well as the baseline. For our method in this limited-data setting, we use $\lambda = 0.8$ for the CUB-200, Cars-196, and VGG Flowers datasets and $\lambda = 0.6$ for the FGVC-Aircraft datasets. For all datasets we use $\tau = 0.5$ as the value for the temperature hyperparameter.

Results for classification accuracy in this limited-data setting are found in Table 7. Both baseline and CGC models perform better with more samples per class, but our method consistently outperforms the baseline. The magnitude of improvement varies from 1-4% points. Notably, our method outperforms the baseline by 4% points (5-shot and 10-shot) on CUB-200 and by 2% points (5-shot and 10-shot) on Cars-196. Since only CUB-200 and Cars-196 contain bounding box annotations, we report the CH evaluation metric for these two datasets in Table 8. CH results are bet-

Method	CUB-200	FGVC-Aircraft	Cars-196	VGG Flowers-102
Baseline	80.09 ± 0.89	83.65 ± 0.15	89.71 ± 0.14	$\mathbf{96.09 \pm 0.23}$
Ours (CGC)	$\mathbf{81.49 \pm 0.09}$	$\mathbf{85.72 \pm 0.20}$	$\mathbf{90.28 \pm 0.08}$	$\mathbf{96.18 \pm 0.09}$

Table 6. Evaluation of classification accuracy of the baseline against our CGC method for fine-grained datasets. We run 3 trials and report the mean and standard deviation. We observe consistent improvements in the classification accuracy across all four datasets with the largest gain on the FGVC-Aircraft dataset.

	CUB200			Cars196		
Method	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Baseline	13.7 ± 0.3	51.7 ± 0.3	66.4 ± 0.2	6.1 ± 0.2	34.3 ± 0.4	61.1 ± 0.4
Ours (CGC)	$\mathbf{15.8 \pm 0.3}$	$\mathbf{55.2 \pm 0.3}$	$\mathbf{68.4 \pm 0.3}$	$\mathbf{6.5 \pm 0.2}$	$\mathbf{36.9 \pm 0.4}$	$\mathbf{63.0 \pm 0.4}$
	Aircrafts			Flowers		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Baseline	7.7 ± 0.3	25.7 ± 0.4	41.4 ± 0.3	52.1 ± 0.5	$\mathbf{85.6 \pm 0.4}$	$\mathbf{93.2 \pm 0.2}$
Ours (CGC)	$\mathbf{8.0 \pm 0.3}$	$\mathbf{26.9 \pm 0.4}$	$\mathbf{42.9 \pm 0.3}$	$\mathbf{53.3 \pm 0.5}$	$\mathbf{85.8 \pm 0.4}$	$\mathbf{93.4 \pm 0.2}$

Table 7. Classification accuracy for the limited-data setting on the fine-grained datasets. We run 20 trials and report the mean and standard deviation. We observe consistent improvement across all datasets with the largest improvements observed on the CUB-200 and Cars-196 datasets.

	CUB-200				Cars-196			
Method	1-shot	5-shot	10-shot	Full training set	1-shot	5-shot	10-shot	Full training set
Baseline	55.54	57.60	61.39	63.71	**60.51**	**64.08**	64.31	65.58
Ours (CGC)	**61.63**	**63.55**	**73.86**	**71.08**	60.48	62.13	**64.63**	**69.04**

Table 8. Content Heatmap (CH) evaluation results for limited and full training data settings on CUB-200 and Cars-196 datasets. CH results are better for all CUB-200 sets but are marginally worse for 1-shot and 5-shot settings on Cars-196. We believe the lower accuracy of the model for few-shot setting results in noisy interpretations, so the CH metric becomes less reliable.

ter for all CUB-200 sets but are marginally worse for 1-shot and 5-shot settings on Cars-196. We believe the lower accuracy of the model for the few-shot setting results in noisy interpretations, so the CH metric becomes less reliable.

4.3. Using Unlabeled Data for CGC Loss

Models trained with categorical cross-entropy loss only have no way to incorporate information from unlabeled data. Since our method does not rely on labels for computing the explanation heatmaps, we extend our method to use mostly unlabeled data with just a small fraction of labeled data for training. We use 1% of ImageNet training images as labeled data and leverage the rest of ImageNet training images as the unlabeled data for the CGC loss term. We initialize both the baseline and our method from a ResNet50 model trained using SwAV [2].

Table 9 shows that our method outperforms the baseline by 1% point on both the Top-1 and Top-5 accuracy metrics. Since we are using only 1% labeled data, the resulting Grad-CAM heatmap will be less accurate and hence the 1% point improvement is not insignificant. Even though our model is only able to leverage the additional unlabeled data for the CGC loss term, and not the categorical accuracy, we can see that the CGC loss acts as a regularizer to improve the generalization of the model.

Model	Top-1 (%)	Top-5 (%)	CH (%)
Baseline	54.00	78.69	46.08
Ours (CGC)	**55.18**	**79.12**	**46.76**

Table 9. On the 1% ImageNet limited-label setting, our CGC method leverages unlabeled data for the L_{CGC} loss term and is able to improve the classification accuracy and explanations when evaluated on ImageNet validation data. Both models are initialized from a ResNet50 model trained in an unsupervised manner using SwaV [2].

4.4. Ablation on λ

We perform an ablation experiment to study the sensitivity of our method to the λ hyperparameter. If λ is too low, the cross-entropy term will dominate the optimization and thus the resulting improvement in the explanation consistency will be marginal, whereas if λ is too high, the CGC loss will be applied on noisy heatmaps resulting in lower accuracy and lower explanation consistency. Table 10 shows the Top-1 accuracy and CH scores for different values of λ for ResNet18 on ImageNet dataset. We choose $\lambda = 1.0$ for ResNet18 as the best trade-off between accuracy and CH.

4.5. Qualitative Results

Figure 3 compares the Grad-CAM heatmaps generated by the baseline model against our CGC model on Ima-

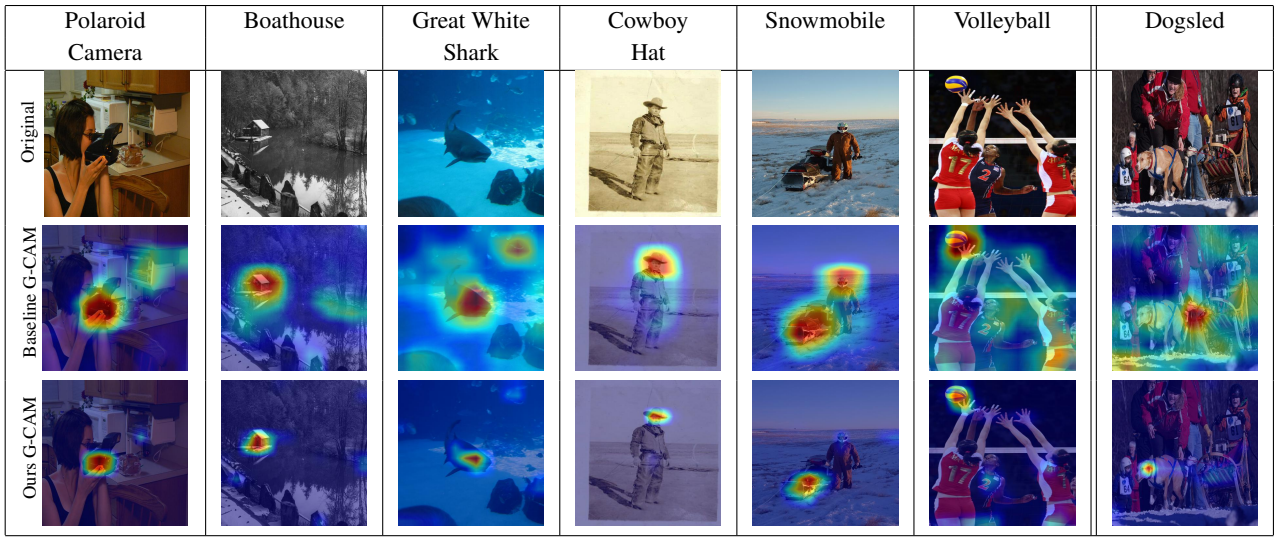

Figure 3. Grad-CAM visualization results for images from ImageNet validation set using ResNet50. We observe that the model trained using our method focuses on the most discriminative regions of the object instead of the background pixels. In the 6th column above, we compute the Grad-CAM explanation for the category "volleyball" and we see that our model no longer focuses on the players and correctly highlights the volleyball as the cause of the target category. The last column contains a failure example for the category "Dogsled". The baseline model correctly highlights the sled whereas our model incorrectly highlights the dog.

λ	0	0.01	0.1	0.5	1.0	2.0
Top-1 Acc	69.76	66.97	66.90	65.14	66.37	65.89
CH	54.47	56.18	61.97	66.19	65.83	52.60

Table 10. Ablation to study the sensitivity of our method to the λ hyperparameter for ResNet18 on ImageNet dataset.

geNet. Our model consistently focuses on the discriminatory parts of the object and does not highlight background pixels. Additional results for CUB-200, Cars-196 and Aircrafts datasets are included in the appendix.

5. Conclusion

We introduce a contrastive learning method for improving the explanations generated by a deep neural network, training them to be consistent with spatial transformations. We emphasize the importance of evaluating the network based on its quality of explanation, and not only classification accuracy. Our CGC method significantly improves the explanation heatmaps while obtaining comparable classification accuracy on ImageNet and UnRel datasets. Furthermore, our method is able to boost the classification accuracy on fine-grained classification datasets such as CUB-200, Cars-196, VGG Flowers-102, and FGVC-Aircraft while improving the consistency of explanation heatmaps with human annotations. This demonstrates that our method acts as a regularizer that focuses more attention on the discriminating aspects of the image. We also show that our method is able to leverage unlabeled data to improve the classification accuracy in limited-label data settings.

Limitations: Our method uses Grad-CAM [32] algorithm to compute the explanation heatmaps for the original set of images as well as the augmented images, which are then used to compute the contrastive loss term L_{CGC}. In comparison to standard training with cross-entropy loss, our method requires additional compute to account for storing the additional gradient graph in memory during backpropagation. While this is an overhead during the training stage, we believe the incurred compute cost is offset by the improved explainability of the resulting model.

Ethics Statement: Our method improves the explainability of image classification models and thereby increases trust and transparency of the underlying decision making process. However, our method is a data-driven approach and hence could reflect potential negative biases present in the training dataset. Moreover, explanation methods such as Grad-CAM [32] can be an unreliable estimate of model interpretability (i.e., evidence for an incorrect prediction looking identical to evidence for a correct prediction).

Acknowledgment: This material is based upon work partially supported by the U.S. Air Force under Contract No. FA8750-19-C-0098, U.S. Department of Commerce, National Institute of Standards and Technology under award number 60NANB18D279, NSF grant numbers 1845216 and 1920079, and funding from Northrop Grumman and SAP SE. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force, DARPA, or other funding agencies. We would also like to thank the reviewers for their valuable feedback.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 2020. 7

[3] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 4

[5] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, pages 248–255. IEEE, 2009. 13

[7] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 3

[8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2950–2958, 2019. 3, 14

[9] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 3

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3

[11] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742, 2006. 2

[13] Tao Han, Wei-Wei Tu, and Yu-Feng Li. Explanation consistency training: Facilitating consistency-based semi-supervised learning with interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7639–7646, 2021. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *ArXiv*, abs/1911.05722, 2019. 2, 3, 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[16] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 3

[17] Akshita Jha, Vineeth Rakesh, Jaideep Chandrashekhar, Adithya Samavedhi, and Chandan K Reddy. Supervised contrastive learning for interpretable long document comparison. *arXiv preprint arXiv:2108.09190*, 2021. 3

[18] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017. 2

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 6, 13

[20] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 840–849, 2017. 3

[21] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 6, 13

[22] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 6, 13

[23] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representions by solving jigsaw puzzles. In *ECCV*, 2016. 3

[24] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5899–5907, 2017. 2, 3

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[26] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models.

In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 3, 5, 14

[27] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017. 5, 6

[28] Vipin Pillai and Hamed Pirsiavash. Explainable models with consistent interpretations. *AAAI*, 2021. 2, 3, 4, 5, 6

[29] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR, 2020. 3

[30] 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. 3

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 5

[32] 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2016. 1, 3, 8

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. 3

[34] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020. 1

[35] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014. 3

[36] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[37] Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[38] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 3

[39] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 1

[40] Dong Wang, Yuewei Yang, Chenyang Tao, Zhe Gan, Liqun Chen, Fanjie Kong, Ricardo Henao, and Lawrence Carin. Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models. *arXiv preprint arXiv:2012.03369*, 2020. 3

[41] Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N. Metaxas. Sharpen focus: Learning with attention separability and consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[42] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 2, 6, 13

[43] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *ArXiv*, abs/1311.2901, 2013. 3

[44] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 5, 6

[45] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016. 3

[46] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2015. 1, 3

6. Appendix

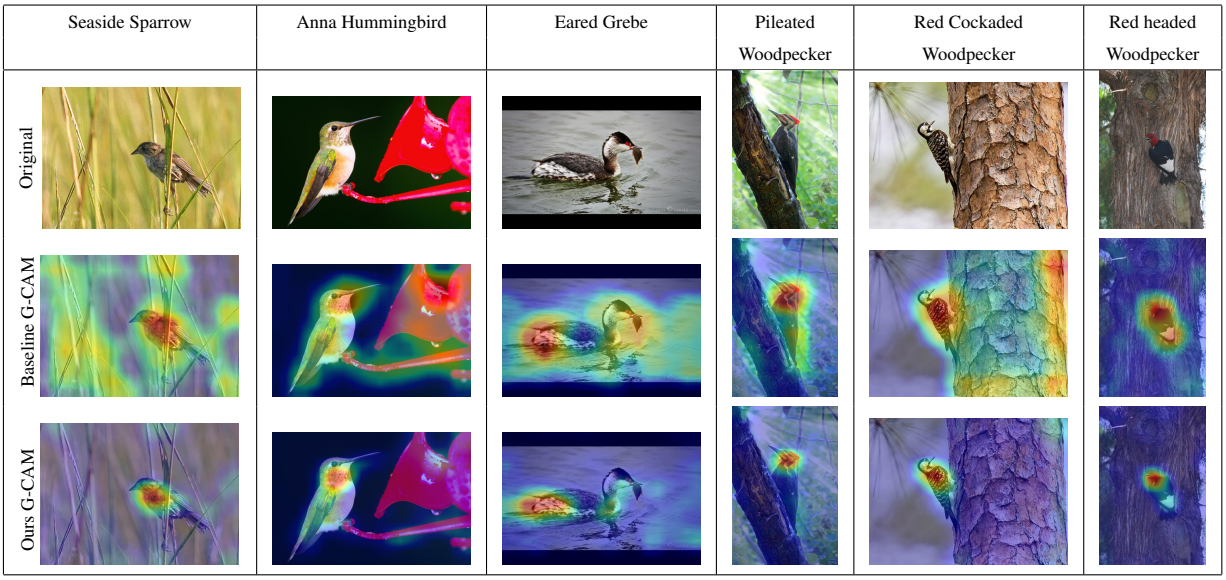

Figure 4. Grad-CAM visualization results for images from the CUB-200 validation set using ResNet50. Interestingly, on the right column, the baseline is focusing on the whole bird while our method focuses on the head of the bird only. Given the name of the category "Red headed Woodpecker", it makes sense that the head should be the most discriminative region.

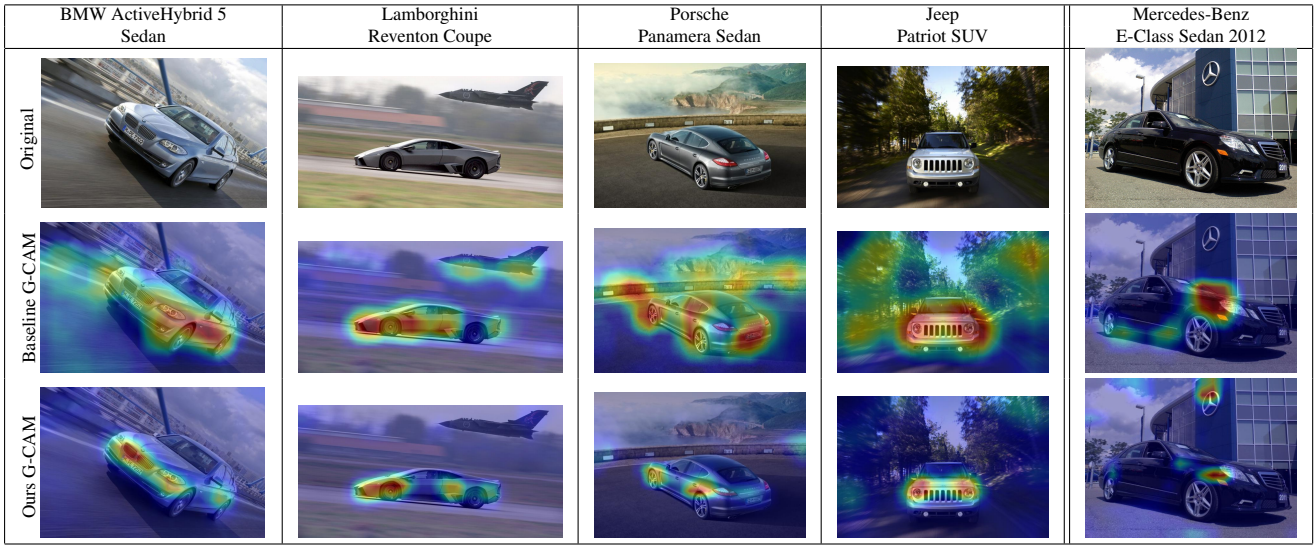

Figure 5. Grad-CAM visualization results for images from the Cars-196 validation set using ResNet50. In the second column above, we see that our model is able to correctly focus on the object regions of "Lamborghini" whereas the baseline incorrectly highlights the fighter jet as well. The last column shows a failure example where our model incorrectly focuses on the "Mercedes" logo on the building instead of the car itself. This is an interesting failure case since the logo is a valid discriminatory attribute for a fine-grained car classification.

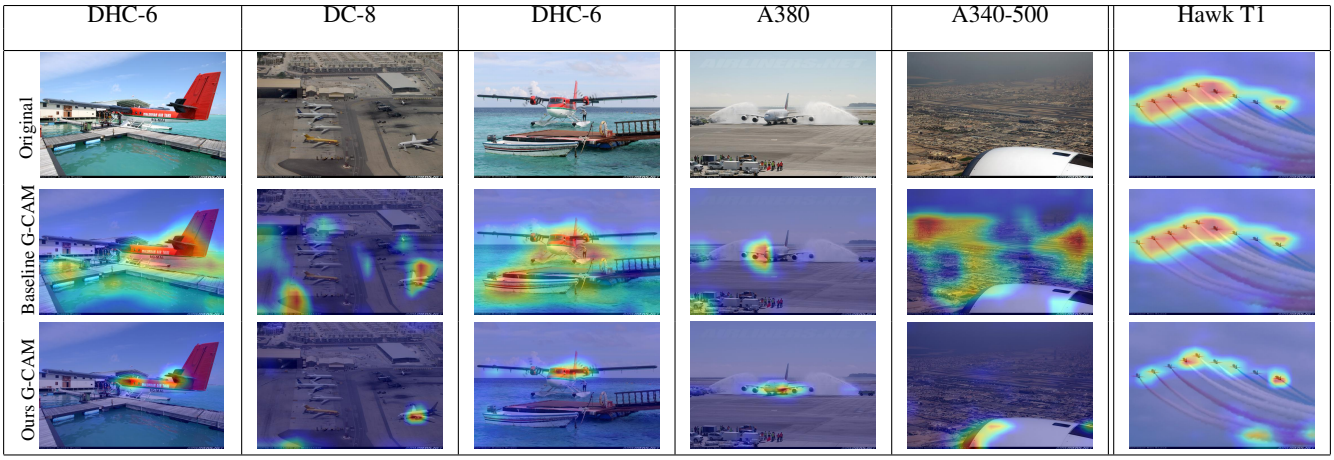

Figure 6. Additional Grad-CAM visualization results on the FGVC Aircraft validation set using ResNet50. While our method correctly highlights most discriminative part of the aircraft in the first and the third column, the baseline incorrectly highlights the water along with the aircraft. Note that the last column shows a failure case for our model which incorrectly highlights the smoke trajectory of the aircraft.

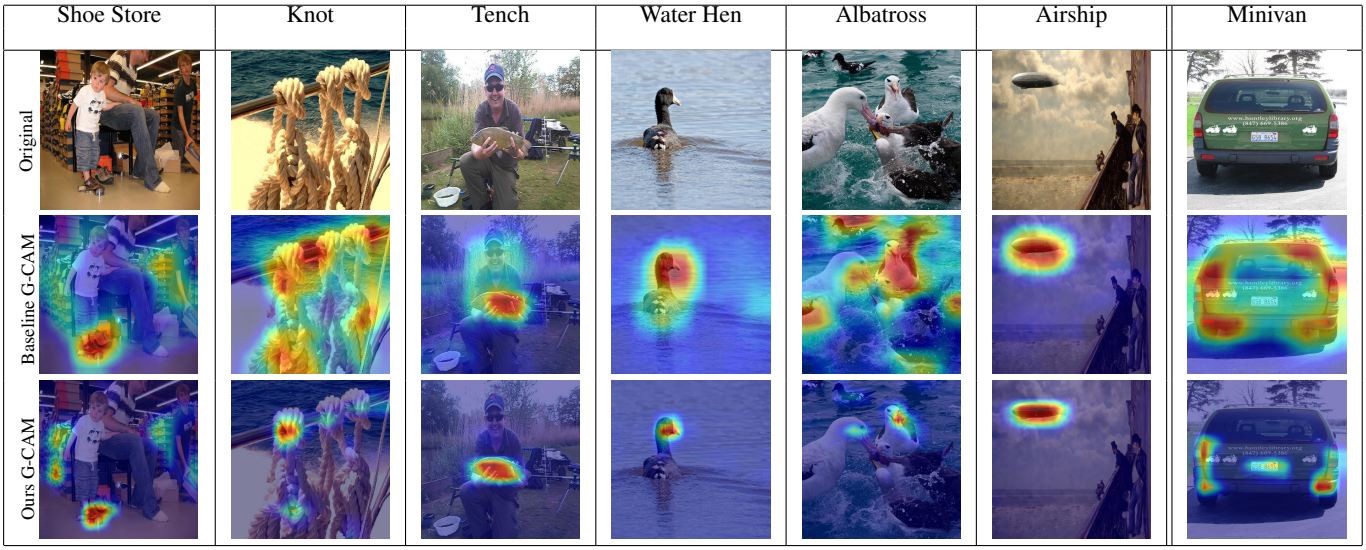

Figure 7. Additional Grad-CAM visualization results on the ImageNet validation set using ResNet50. Our model is able to improve upon the baseline by not relying on background pixels and instead focusing on the most discriminative regions of the object. In the 3rd column, Grad-CAM is computed for the category "Tench"and we see that the baseline incorrectly highlights the person along with the fish whereas our model correctly highlights the fish. The last column shows a failure case where our model incorrectly highlights the license plate along with the other regions of the minivan.

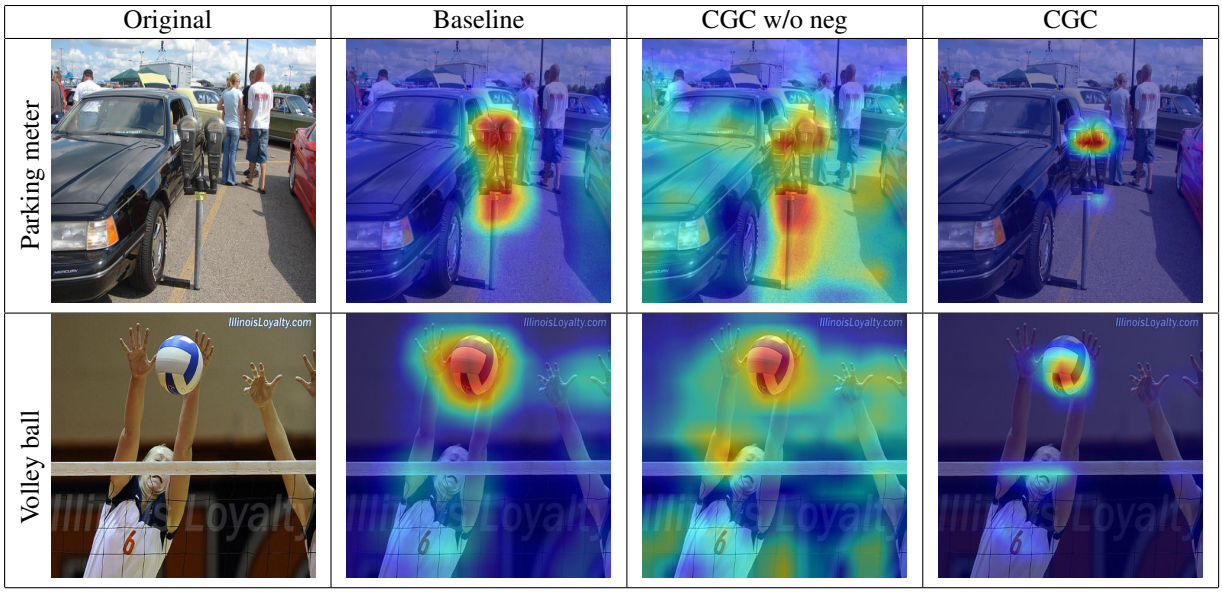

Figure 8. Qualitative results for Table 5 in the main paper. The Grad-CAM heatmap for the model trained with CGC loss, but without the negative examples results in a uniform heatmap spread across the image (column 3). In the second row, the baseline model relies on the arms of the player for classifying the image as 'Volleyball', whereas our method is able to reduce this spurious correlation.

Model	ResNet50 ImageNet-100	
	Top-1 Acc (%)	CH (%)
Baseline	**86.40**	53.60
CGC	84.04	**72.32**
CGC w/o neg	81.94	38.46

Table 11. Results similar to Table 5 of the main paper with ResNet50 on ImageNet-100 (subset of ImageNet with 100 classes). The low CH for CGC w/o negatives shows that this method could result in heatmaps diffused across the image. The model trained without the negative heatmaps as part of L_{CGC} loss results in a very low CH, thus confirming our hypothesis that the lack of negative heatmaps would result in a model learning a trivial solution of generating heatmaps diffused throughout the image.

7. License for assets

We list the license for each of the dataset and code assets used for our experiments.

ImageNet: We have been granted access to the ImageNet [6] dataset for non-commercial research/educational purposes and we abide by the terms of the license of this dataset.

CUB-200: This dataset was introduced in [42] and we use the images and the accompanying annotations for non-commercial/education research purposes only.

FGVC-Aircraft: The images in the FGVC-Aircraft [21] dataset has been made available exclusively for non-commercial research/educational purposes and as such we only use this dataset for non-commercial research/educational purposes.

Stanford Cars-196: This dataset [19] has been made available for non-commercial research purposes only and we abide by the terms of the license of this dataset.

VGG Flowers-102: The images and annotations in the VGG Flowers-102 [22] dataset are released under the GNU General Public License, version 2.

TorchRay: This framework was introduced in [8] and is licensed under CC-BY-NC. We use this framework for evaluating the explanation heatmaps using the Content Heatmap (CH) metric.

Insertion AUC metric: This metric was introduced in [26] and the accompanying code is released under the MIT license.