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

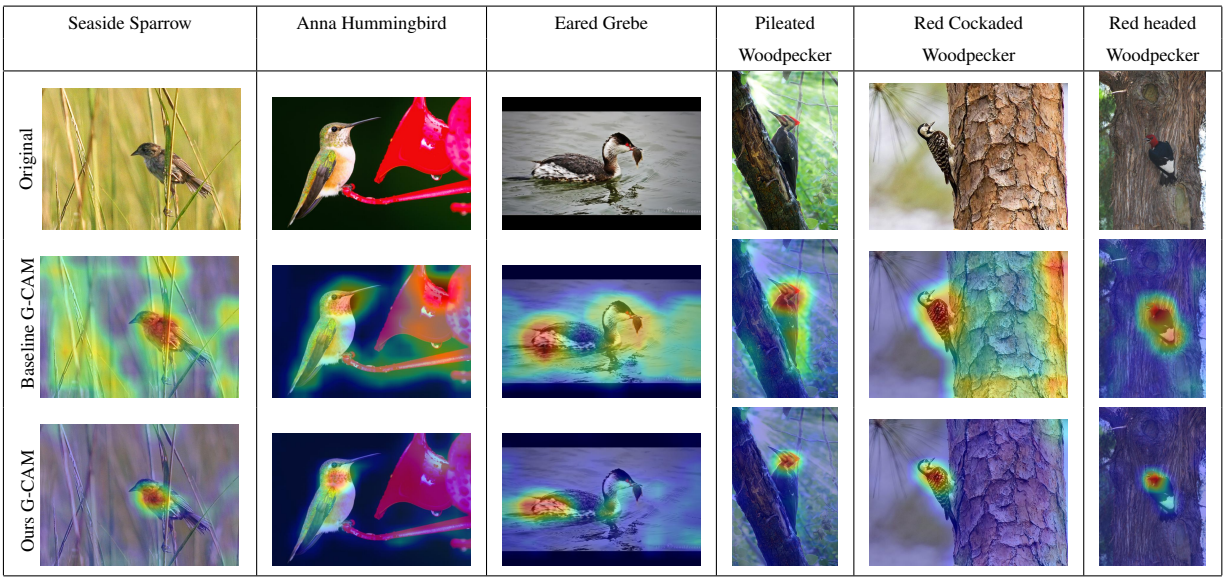

Figure 1. Grad-CAM visualization results for images from the CUB-200 validation set using ResNet50. Interestingly, on the right column, the baseline is focusing on the whole bird while our method focuses on the head of the bird only. Given the name of the category "Red headed Woodpecker", it makes sense that the head should be the most discriminative region.

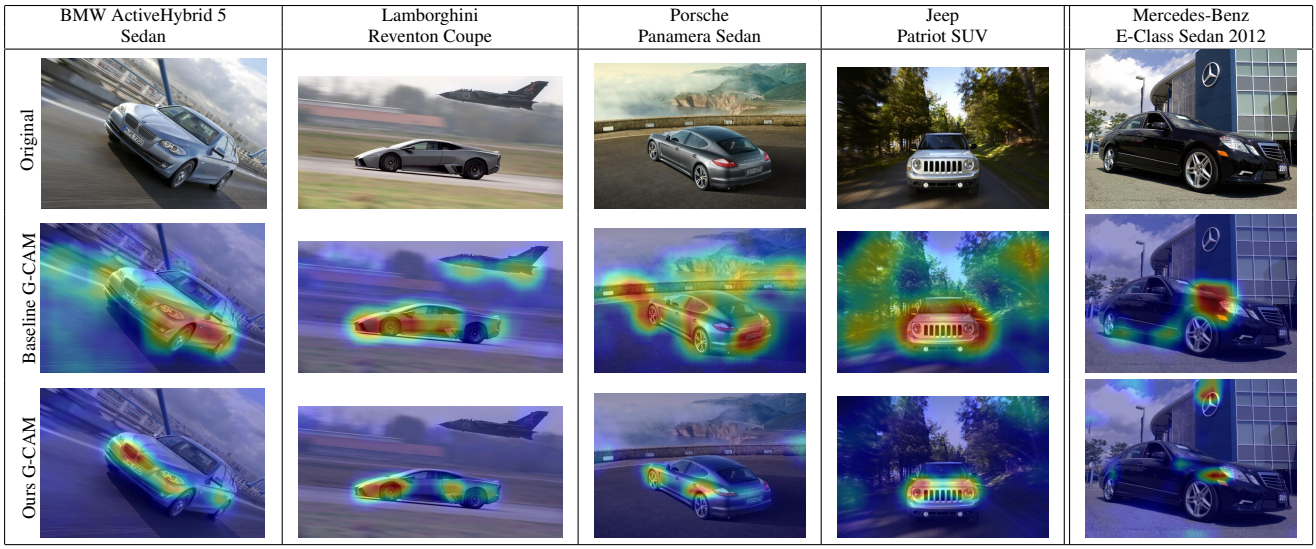

Figure 2. Grad-CAM visualization results for images from the Cars-196 validation set using ResNet50. In the second column above, we see that our model is able to correctly focus on the object regions of "Lamborghini" whereas the baseline incorrectly highlights the fighter jet as well. The last column 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.

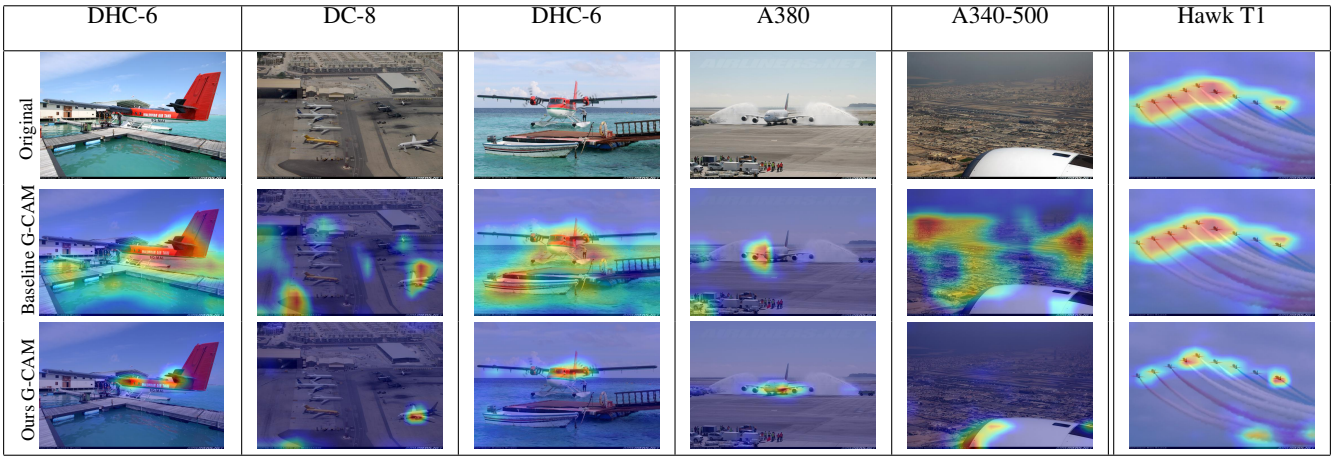

Figure 3. Additional Grad-CAM visualization results on the FGVC Aircraft validation set using ResNet50. While our method correctly highlights most discriminative part of the aircraft in the first and the third column, the baseline incorrectly highlights the water along with the aircraft. Note that the last column shows a failure case for our model which incorrectly highlights the smoke trajectory of the aircraft.

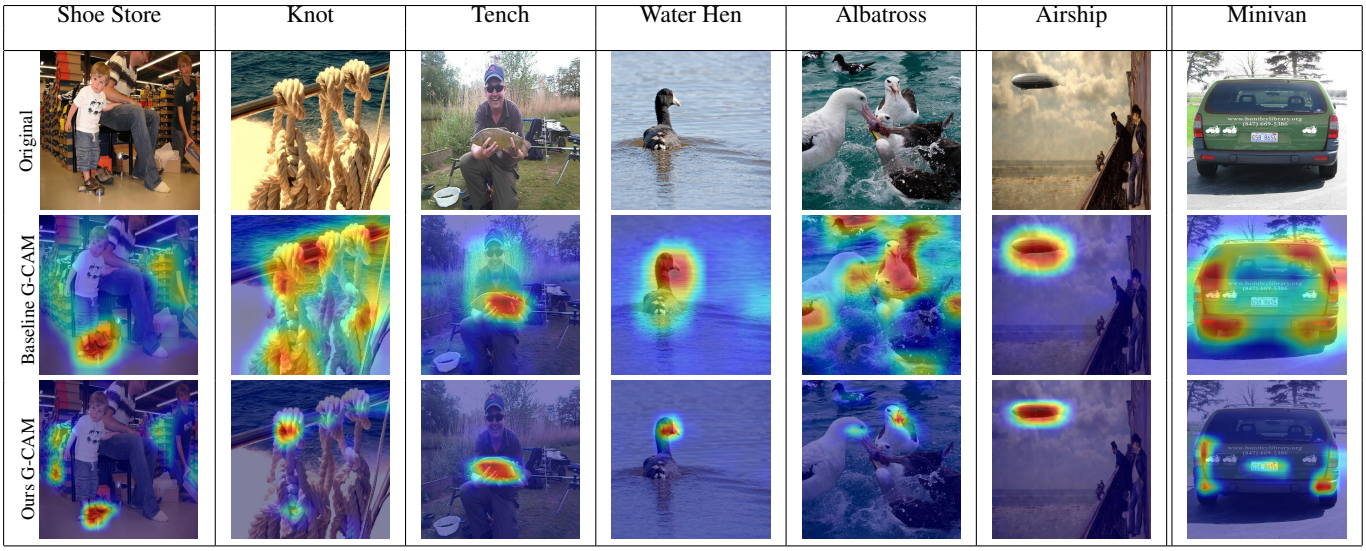

Figure 4. Additional Grad-CAM visualization results on the ImageNet validation set using ResNet50. Our model is able to improve upon the baseline by not relying on background pixels and instead focusing on the most discriminative regions of the object. In the 3rd column, Grad-CAM is computed for the 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.

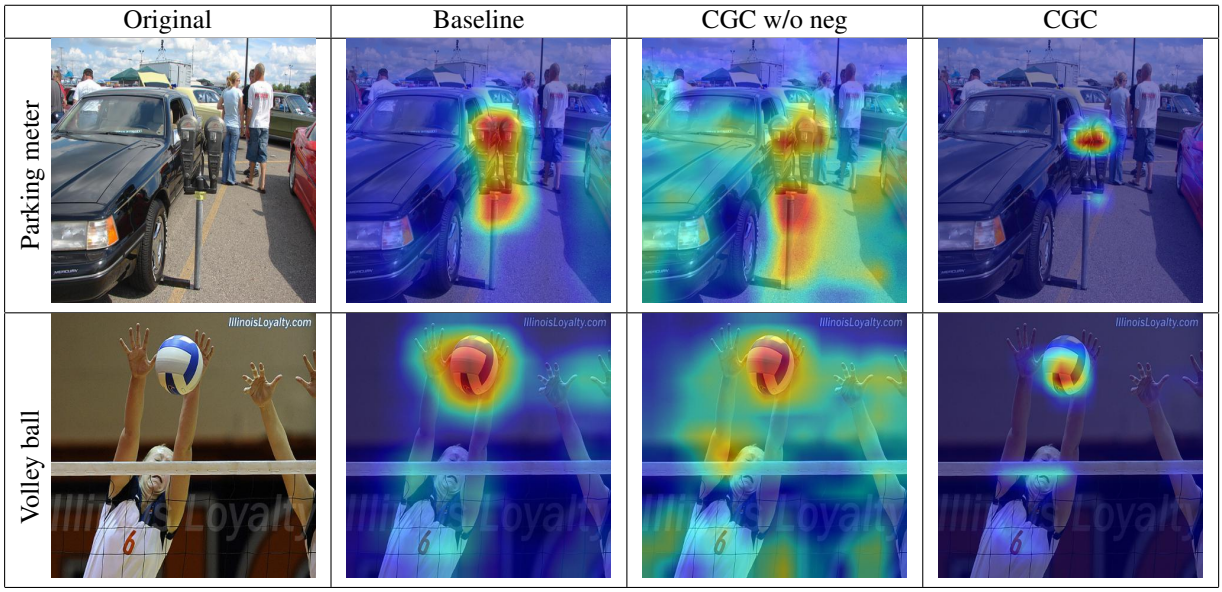

Figure 5. Qualitative results for Table 5 in the main paper. The Grad-CAM heatmap for the model trained with CGC loss, but without the negative examples results in a uniform heatmap spread across the image (column 3).

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 1. 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.

1. 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 [1] 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 [7] and we use the images and the accompanying annotations for non-commercial/education research purposes only.

FGVC-Aircraft: The images in the FGVC-Aircraft [4] 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 [3] 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 [5] dataset are released under the GNU General Public License, version 2.

TorchRay: This framework was introduced in [2] 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 [6] and the accompanying code is released under the MIT license.

References

[1] 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. 4

[2] 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. 4

[3] 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. 4

[4] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 4

[5] 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. 4

[6] 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. 4

[7] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 4