Consistent Explanations by Contrastive Learning

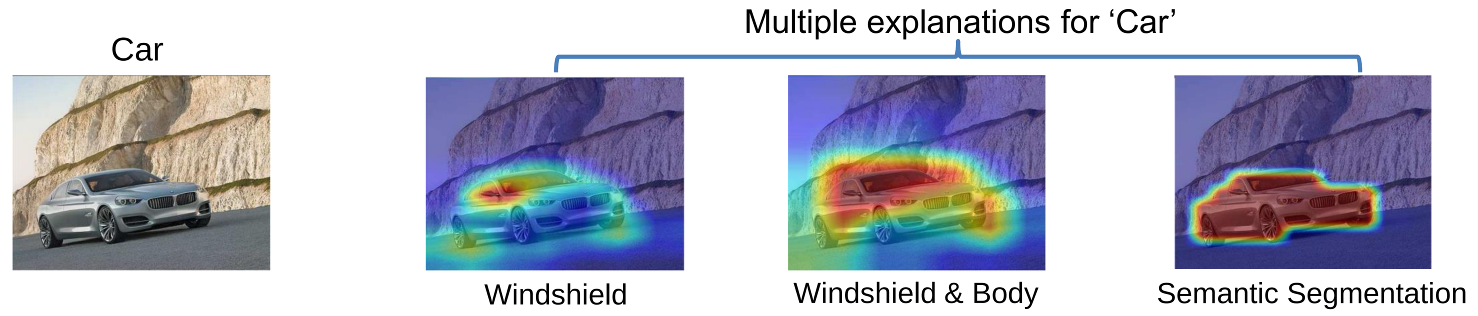
Vipin Pillai[1], Soroush Abbasi Koohpayegani[1], Ashley Ouligian[2], Dennis Fong[2], Hamed Pirsiavash[1,3]
[1]University of Maryland, Baltimore County [2]Northrop Grumman [3]University of California, Davis

PyTorch Code

Motivation

- Our goal is to train models that are more explainable given an explanation tool (e.g., Grad-CAM).
- Image classifiers may learn unwanted contextual biases from data

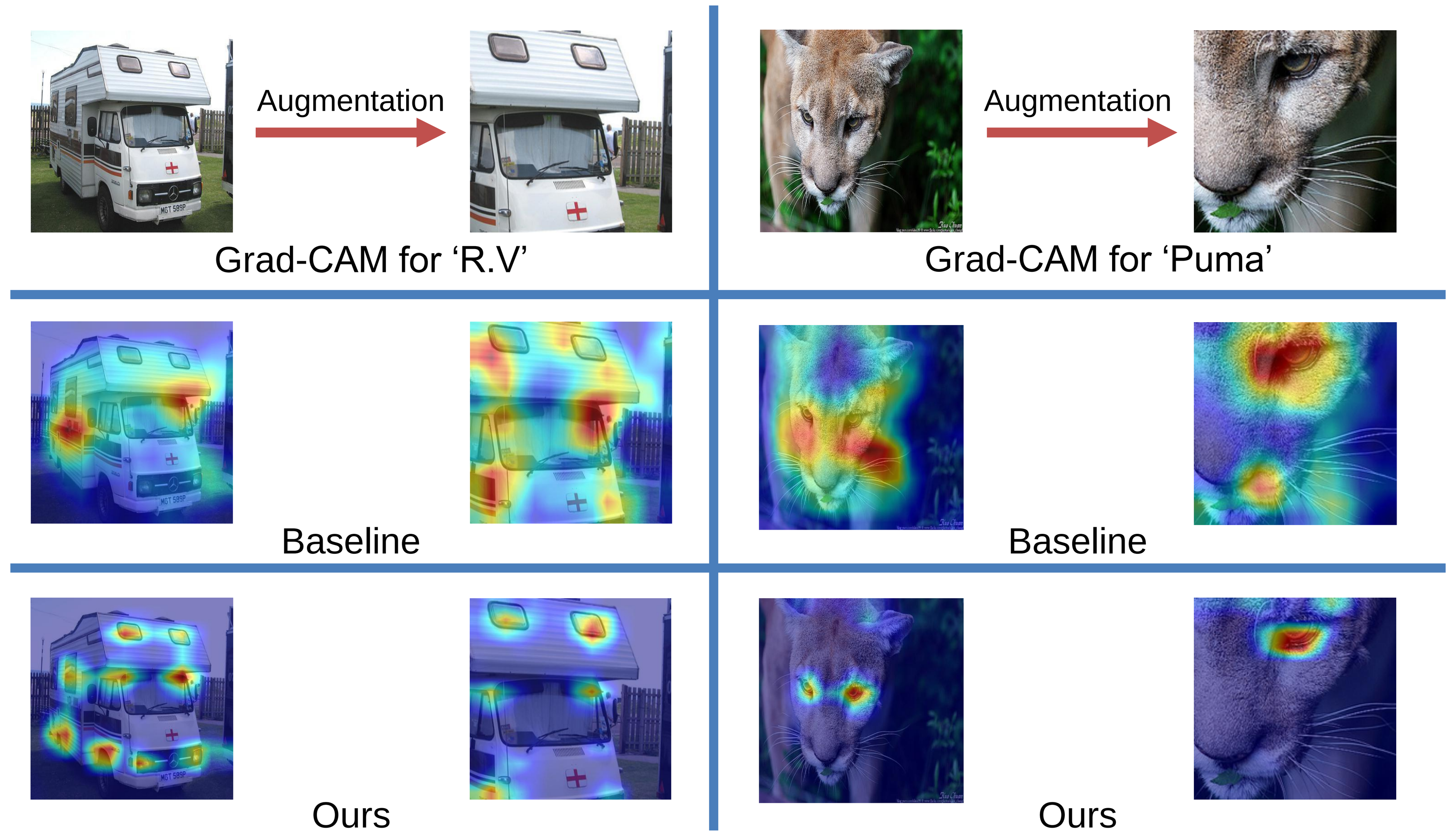
- Supervising explanation is not feasible since annotating explanations is not a well-defined task.



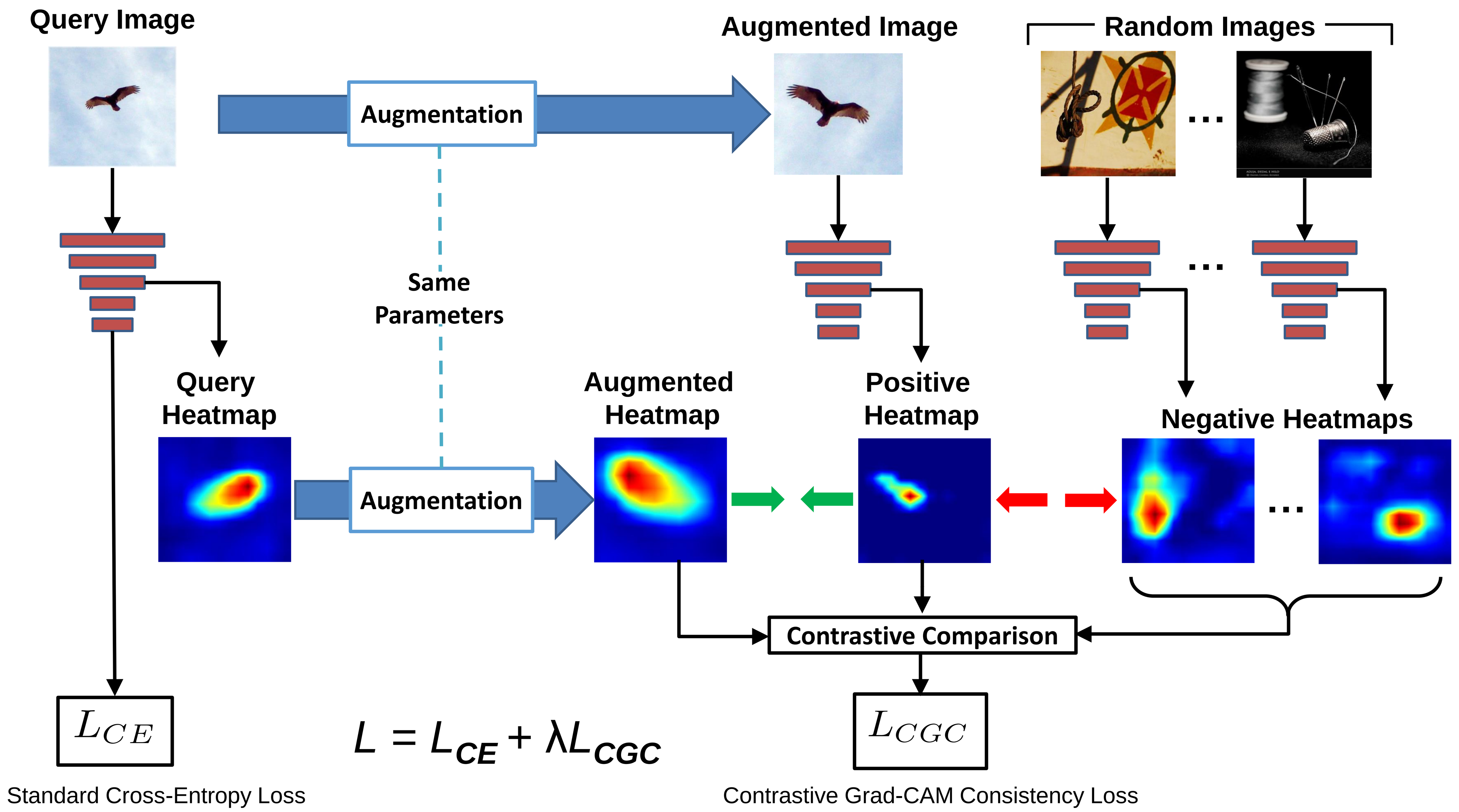
- We adopt ideas from self-supervised learning to improve explanation without annotating explanation.
- Our method can also leverage unlabeled data.

Key Idea

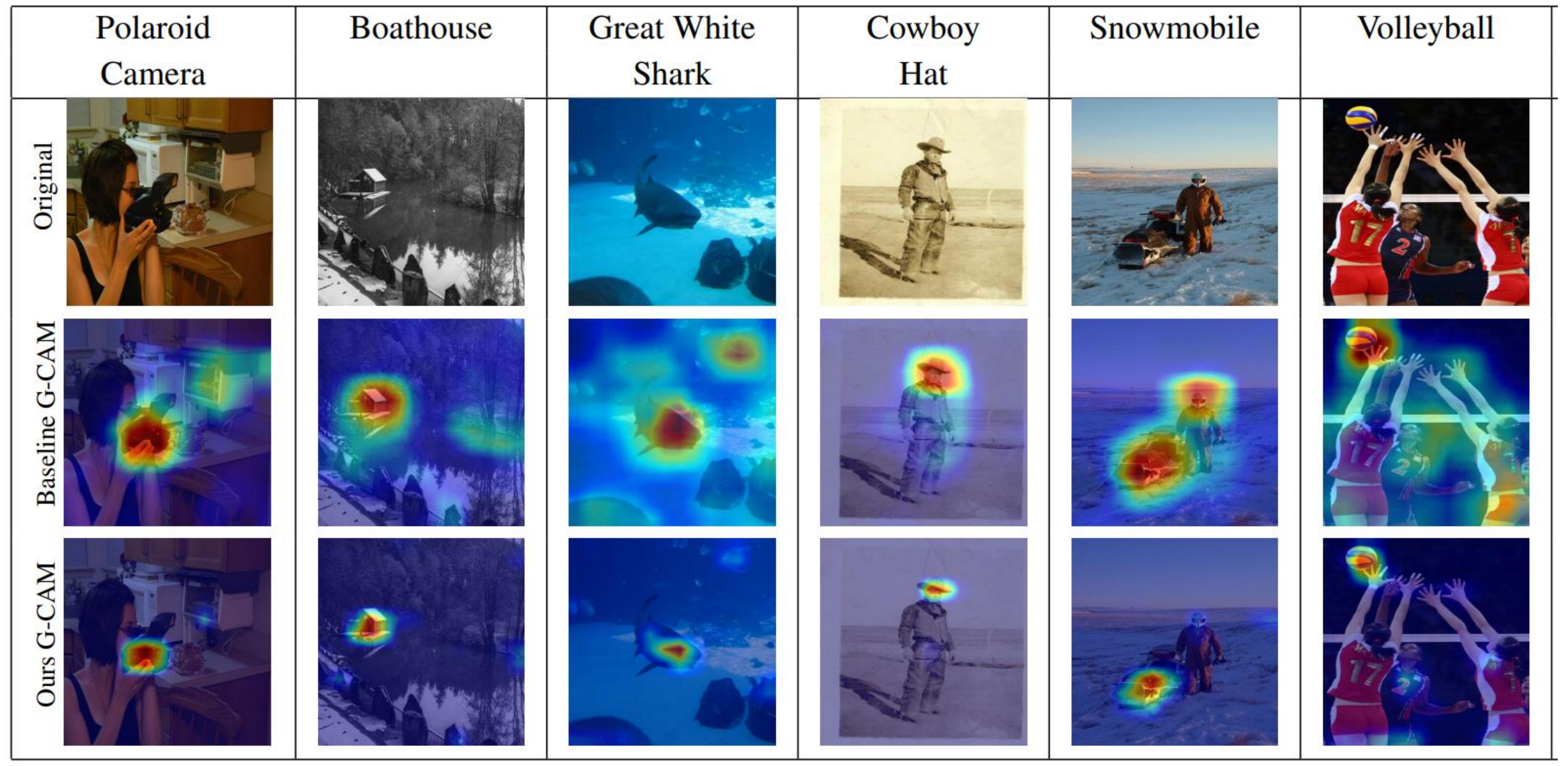
- Consistent explanations w.r.t. spatial transformations



Method: Contrastive learning on the explanation space



Results



Our model focuses on the most discriminative regions of the object instead of the background pixels.

Content Heatmap (CH) [1]: Percentage of heatmap strictly within object bounding box.

ImageNet

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

Our method results in models with improved explanation with marginal drop in classification accuracy.

Fine-grained Classification

Method	CUB-200	FGVC-Aircraft	Cars-196	VGG Flowers-102
Cross-Entropy	80.09 ± 0.89	83.65 ± 0.15	89.71 ± 0.14	**96.09 ± 0.23**
Ours (CGC)	**81.49 ± 0.09**	**85.72 ± 0.20**	**90.28 ± 0.08**	**96.18 ± 0.09**

Our method improves the classification accuracy on fine-grained datasets.

1% Labeled ImageNet Subset

Method	Top-1 Acc (%)	Top-5 Acc (%)	CH (%)
Cross-Entropy	54.00	78.69	46.08
Ours (CGC)	**55.18**	**79.12**	**46.76**

Our method leverages unlabeled data for L_{CGC} loss term (both models are initialized from SwaV[2])

Conclusion

- We introduce a contrastive learning method for training **more explainable** models for a given explanation tool (e.g., Grad-CAM).
- Compared to manual annotation, our method significantly improves explanation consistency on ImageNet, UnRel, and fine-grained datasets.
- Our method acts as a regularizer that focuses more attention on the discriminating aspects of the image, thereby **reducing contextual bias**.
- We further leverage **unlabeled data** to improve the classification accuracy in **limited-label settings**.

References

[1] Vipin Pillai and Hamed Pirsiavash. Explainable models with consistent interpretations. AAAI 2021.
[2] Caron, Misra, Mairal, Goyal, Bojanowski, Joulin. Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS 2020.