Are Data-driven Explanations Robust against Out-of-distribution Data? Supplementary Material

Tang Li Fengchun Qiao Mengmeng Ma Xi Peng
University of Delaware
{tangli, fengchun, mengma, xipeng}@udel.edu

1. Additional Results

1.1. Robustness against Out-of-distribution Data

We empirically investigate the robustness of data-driven explanations against *out-of-distribution* data. We evaluate the problem on an OOD generalization benchmark image dataset *Terra Incognita* [2] and a scientific tabular dataset *Urban Land* [5]. For image data, we evaluate ERM [18], GroupDRO [14] and IRM [1] methods, and leverage the Grad-CAM [15] to generate explanations. For scientific tabular data, we use the Input Gradient [16] method to generate explanations, and leverage scientific consistency as the metric to measure explanation quality. Tab. 1 and Tab. 2 show a more detailed comparison of each specified distribution as the testing set. For image data, we can observe a constant explanation fidelity drop on the out-of-distribution data for all tested methods. For scientific tabular data, we can observe a constant scientific consistency drop on the OOD data for all continental regions.

1.2. Evaluation on Terra Incognita

In the experiment on *Terra Incognita* [2], additional visual comparisons on class *cat* and *coyote* are shown in Fig. 1. Results demonstrate that our method can alleviate the model's reliance on *spurious correlations* (*e.g.*, background pixels), and makes consistent explanations on the *out-of-distribution* data.

1.3. Evaluation on VLCS

In the experiment on *VLCS* [4], additional visual comparisons on class *dog* and *person* are shown in Fig. 2. Results demonstrate that our method can alleviate the model's reliance on *spurious correlations* (*e.g.*, background pixels), and makes consistent explanations on the *out-of-distribution* data. Note that our explanations depict the contour of the object better than other explanations.

1.4. Generalize to Different Explanation Methods

In the experiment on generalizing to different explanation methods, additional visual comparisons on *VLCS* [4] dataset are shown in Fig. 3. Results demonstrate that our model's advanced explainability can be generalized to a variety of data-driven explanation methods, such as Integrated Gradients (IG) [17] and Gradient SHAP [10]. Our method significantly alleviate the model's reliance on *spurious correlations* (*e.g.*, tree branches), and makes consistent explanations on the *out-of-distribution* data. Note that our explanations clearly depict the contour of the object.

Method	Test Distribution	iAUC ↑		Δ ↓
		ID	OOD	
ERM [18]	Location 100	0.761	0.517	**0.244**
	Location 38	0.780	0.644	0.136
	Location 43	0.806	0.614	0.192
	Location 46	0.783	0.560	0.223
	Avg.	0.778	0.584	0.194
GroupDRO [14]	Location 100	0.726	0.687	0.039
	Location 38	0.738	0.578	0.160
	Location 43	0.738	0.608	0.130
	Location 46	0.766	0.525	**0.241**
	Avg.	0.742	0.597	0.145
IRM [1]	Location 100	0.575	0.489	0.086
	Location 38	0.745	0.651	0.094
	Location 43	0.539	0.438	**0.101**
	Location 46	0.589	0.500	0.089
	Avg.	0.612	0.520	0.092

Table 1. Evaluation of the explanation fidelity (iAUC) of *in-distribution* (ID) and *out-of-distribution* (OOD) data on *Terra Incognita* [2] dataset. Each specified distribution serves as the testing set. Note that the explanation fidelity severely dropped on OOD data for all tested methods. Specifically, although the IRM method performs a fewer explanation fidelity drop, its in-distribution iAUC is much lower than other methods.

2. Experimental Details

2.1. Architecture Design and Hyper-parameters

We provide additional experimental details on the two image datasets and one scientific tabular dataset: *Terra*

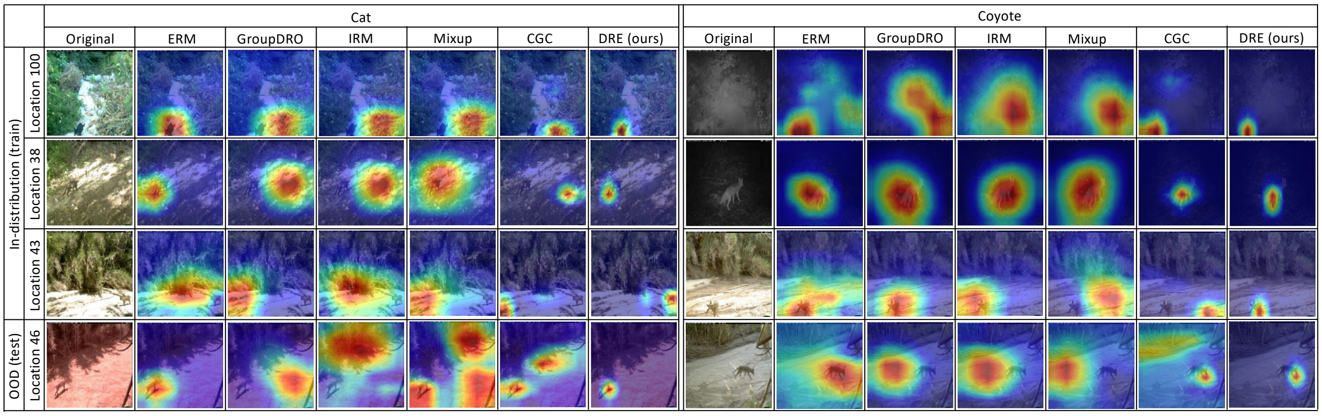

Figure 1. Grad-CAM explanations for images from *Cat* (left) and *Coyote* (right) classes in *Terra Incognita* [2] dataset. The model trained via existing methods, such as ERM [18], Mixup [19], and CGC [11], not only focuses on the objects, but also distribution-specific associations, it getting even severe on *out-of-distribution* data. On the contrary, our model alleviates the reliance on *spurious correlations* (*e.g.*, background pixels), and makes consistent explanations on OOD data. This figure is best viewed in color.

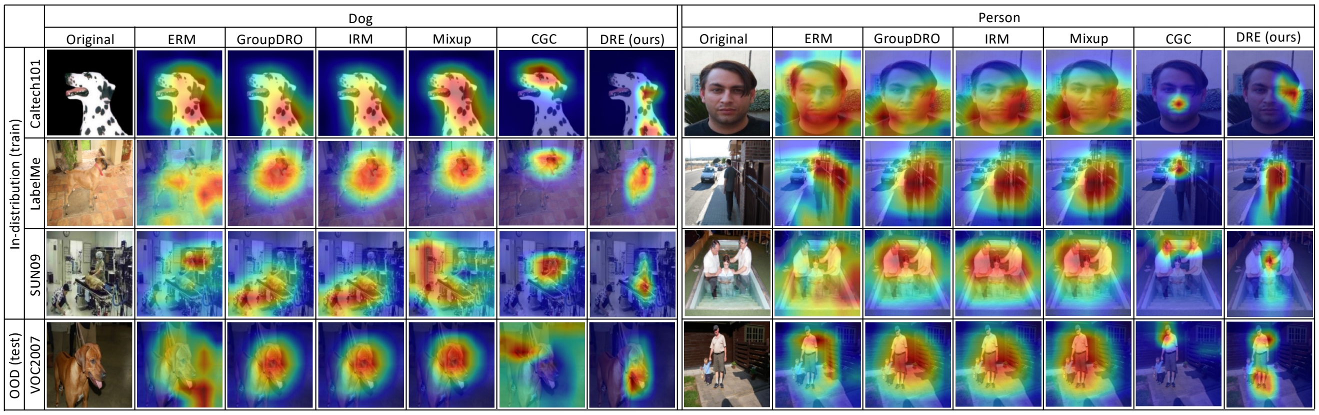

Figure 2. Grad-CAM explanations for images from *Dog* (left) and *Person* (right) classes in *VLCS* [4] dataset. The model trained via existing methods, such as ERM [18], Mixup [19], and CGC [11], not only focuses on the objects, but also distribution-specific associations, it getting even severe on *out-of-distribution* data. On the contrary, our model alleviates the reliance on *spurious correlations* (*e.g.*, background pixels), and makes consistent explanations on OOD data. Note that our explanations depict the contour of the object better than other explanations.

Incognita [2], *VLCS* [4], and *Urban Land* [5]. Following the settings in [6] and [9], Tab. 5 lists all hyperparameters, their default values and random search distribution. We optimize all models using Adam [8].

For *Terra Incognita* [2] and *VLCS* [4], the backbone model is “ResNet-50” [7], detailed architecture is shown in Tab. 3.

For *Urban Land* [5], the backbone model is “U-Net” [12], detailed architecture is shown in Tab. 4.

2.2. Employed Gradient-based Methods

We provide additional details of the gradient-based explanation methods employed in the proposed *Distributionally Robust Explanations* (DRE) method. Built upon our discussions in the Related Work section of the main paper, gradient-based explanation methods offer two properties: (i) computation efficiency; (ii) fully differentiable and thus can be integrated into optimization.

For scientific tabular data (*Urban Land* [5]), we leverage the Input Gradient [16] to calculate explanations, because of its fine-grained resolution and advanced explanation performance on truly continuous inputs [13]. Denote x as an input sample, the predictive model provides a scalar logit $f(x)$ for a particular prediction. An explanation method $g(\cdot)$ calculates an explanation (*e.g.*, heatmap) with the same size as the input, which attributes the model's decision to the features with a higher score. The Input Gradient corresponds to the gradient of the scalar logit for a prediction

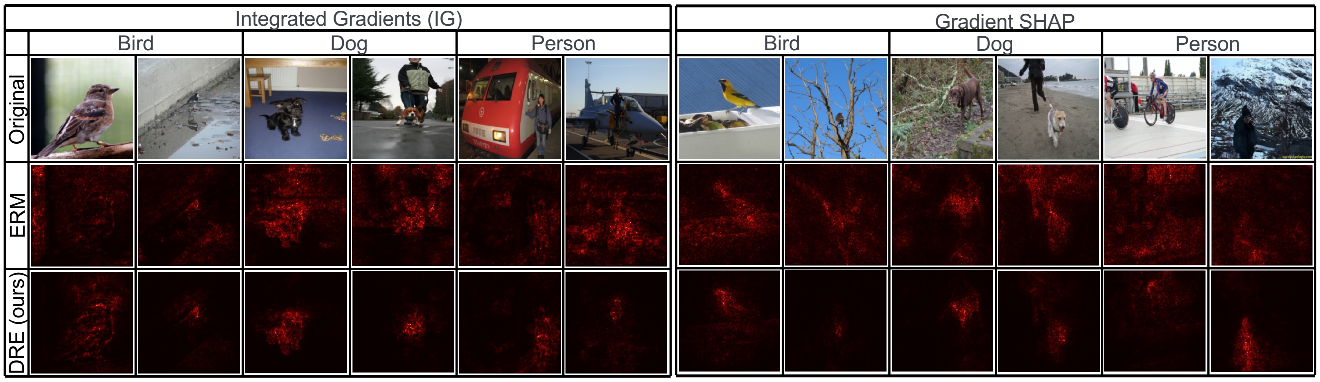

Figure 3. Integrated Gradients (IG) [17] and Gradient SHAP [10] saliency maps for *out-of-distribution* data from *VLCS* [4] dataset. Using VOC2007 [3] as the testing set, for ERM model the saliency maps of both Integrated Gradients (IG) [17] and Gradient SHAP [10] methods are excessively focused on background pixels, such as branch and ground. Our explanations significantly alleviate the salience of *spurious correlations* and clearly depict the contour of the object.

Test Distribution	SC ↑		Δ ↓
	ID	OOD	
Africa	0.968	0.810	0.158
E. Asia	0.835	0.779	0.056
Europe	0.169	0.100	0.069
N. Africa	0.545	-0.707	**1.252**
N. America	0.236	-0.653	0.889
Oceania	0.944	0.855	0.089
Russia	0.257	-0.856	1.113
S. America	0.341	-0.837	1.178
S. Asia	0.520	-0.504	1.024
Avg.	0.535	-0.113	0.648

Table 2. Evaluation of the scientific consistency (SC) of *in-distribution* (ID) and *out-of-distribution* (OOD) data on *Urban Land* [5] dataset using the ERM [18] method. Note that the scientific consistency severely dropped on the *out-of-distribution* data for all tested distributions.

with regard to the input, namely:

$$g_{\text{grad}}(x) = \frac{\partial f(x)}{\partial x}. \tag{1}$$

For image data (*Terra Incognita* [2] and *VLCS* [4]), we leverage the Grad-CAM [15] to calculate explanations. Its superior performance has been empirically proved on tasks such as explaining classification results and weakly supervised semantic segmentation. Let A^k as the feature maps of the last convolutional layer of a DNN, the neuron importance weights:

$$\alpha^k = \frac{1}{Z}\textstyle\sum_i\sum_j \frac{\partial f(x)}{\partial A^k_{ij}}. \tag{2}$$

then the Grad-CAM explanation corresponds to the weighted average of feature maps of the last convolutional layer, namely:

$$g_{\text{grad}-\text{cam}}(x) = \text{ReLU}(\textstyle\sum_k \alpha^k A^k) \tag{3}$$

layer name	output size	ResNet50 layers
conv1	112 × 112	7 × 7, 64, stride 2
conv2_x	56 × 56	3 × 3, max pool, stride 2 $\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times 3$
conv3_x	28 × 28	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times 4$
conv4_x	14 × 14	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times 6$
conv5_x	7 × 7	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times 3$
	1 × 1	average pool, 10-d fc, softmax
FLOPs		3.8×10^9

Table 3. Architecture for *Terra Incognita* [2] dataset. For *VLCS* [4] dataset, the only difference is to change the fc-layer to 5-d.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1, 2, 3

layer names	output size	layer (filter size)	#filters
inconv	16 × 16	Conv(3,3) × 2	64
down1	8 × 8	Conv(3,3) × 2 Maxpool(2,2) × 1	128
down2	4 × 4	Conv(3,3) × 2 Maxpool(2,2) × 1	256
down3	2 × 2	Conv(3,3) × 2 Maxpool(2,2) × 1	512
down4	1 × 1	Conv(3,3) × 2 Maxpool(2,2) × 1	1024
up1	2 × 2	Conv(3,3) × 2 Deconv(2,2) × 1	512
up2	4 × 4	Conv(3,3) × 2 Deconv(2,2) × 1	256
up3	8 × 8	Conv(3,3) × 2 Deconv(2,2) × 1	128
up4	16 × 16	Conv(3,3) × 2 Deconv(2,2) × 1	64
outconv	16 × 16	Conv(1,1) × 1	1
FLOPs		1.2×10^7	

Table 4. The U-Net [12] architecture for *Urban Land* [5] dataset.

Condition	Parameter	Default value	Random distribution
ResNet50	learning rate	5e-5	$10^{\mathrm{Uniform}(-5,-3.5)}$
	batch size	32	$2^{\mathrm{Uniform}(3,6)}$
	dropout	0	[0, 0.1, 0.5]
U-Net	learning rate	1e-4	$10^{\mathrm{Uniform}(-5,-1)}$
	batch size	32	$2^{\mathrm{Uniform}(3,6)}$
	dropout	0	[0, 0.1, 0.5]
DRE	lambda	1.0	$10^{\mathrm{Uniform}(-3,1)}$
	gamma	0.1	$10^{\mathrm{Uniform}(-3,1)}$
GroupDRO	eta	0.01	$10^{\mathrm{Uniform}(-1,1)}$
IRM	lambda	100	$10^{\mathrm{Uniform}(-1,5)}$
	warmup iter.	500	$10^{\mathrm{Uniform}(0,4)}$
Mixup	alpha	0.2	$10^{\mathrm{Uniform}(0,4)}$
CGC	lambda	1.0	$10^{\mathrm{Uniform}(-3,1)}$

Table 5. Hyperparameters, default values, and their distributions.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 3

[4] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 1, 2, 3

[5] Jing Gao and Brian C O'Neill. Mapping global urban land for the 21st century with data-driven simulations and shared socioeconomic pathways. *Nature communications*, 11(1):1–12, 2020. 1, 2, 3, 4

[6] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 2

[7] 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. 2

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 2

[9] Tang Li, Jing Gao, and Xi Peng. Deep learning for spatiotemporal modeling of urbanization. *Advances in Neural Information Processing Systems Workshops*, 2021. 2

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 1, 3

[11] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022. 2

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[13] 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 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670, 2017. 2

[14] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 1

[15] 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. 1, 3

[16] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 3

[18] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 1, 2, 3

[19] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020. 2