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

Tang Li Fengchun Qiao Mengmeng Ma Xi Peng
Department of Computer & Information Sciences
University of Delaware

Poster: TUE-AM-364

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

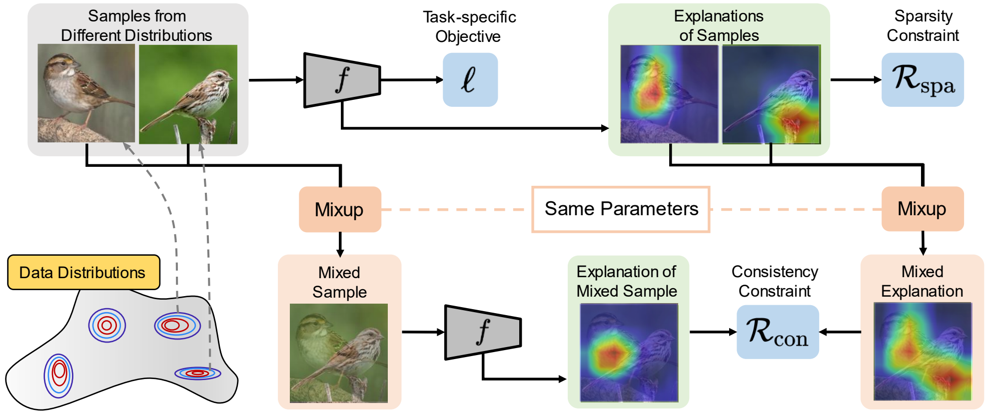
Tang Li, Fengchun Qiao, Mengmeng Ma, Xi Peng
Department of Computer & Information Sciences, University of Delaware

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Highlights

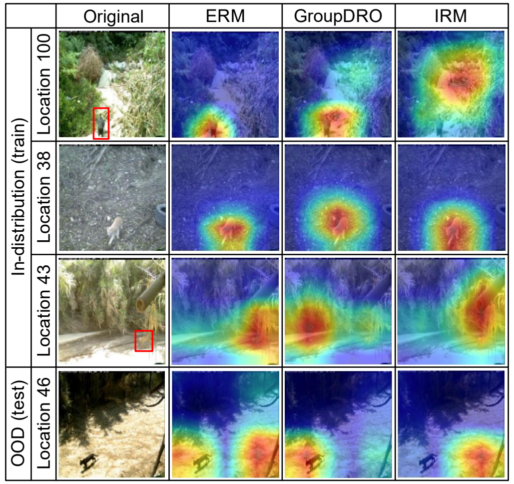
- *Observation*: data-driven explanations are ***unreliable*** on out-of-distribution (OOD) data.
- *Method*: framework **Distributionally Robust Explanations (DRE)** for the learning of consistent explanations across distributions.
- *Experiments*: when testing on OOD data, our model significantly improve the explanation fidelity by ***6.0%*** and prediction accuracy by ***6.9%*** on Terra Incognita dataset.
- Code and pre-trained weights: https://github.com/tangli-udel/DRE

RQ2. *How to develop robust explanations against out-of-distribution data?*

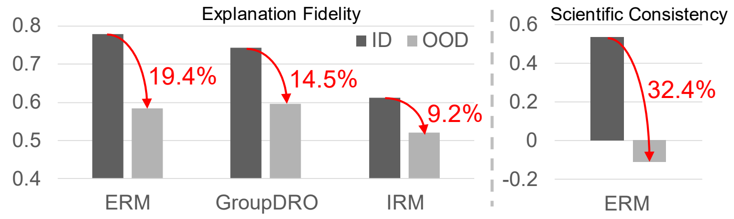


Overview of the proposed Distributionally Robust Explanation (DRE).

RQ1. *Are data-driven explanations robust against out-of-distribution data?*



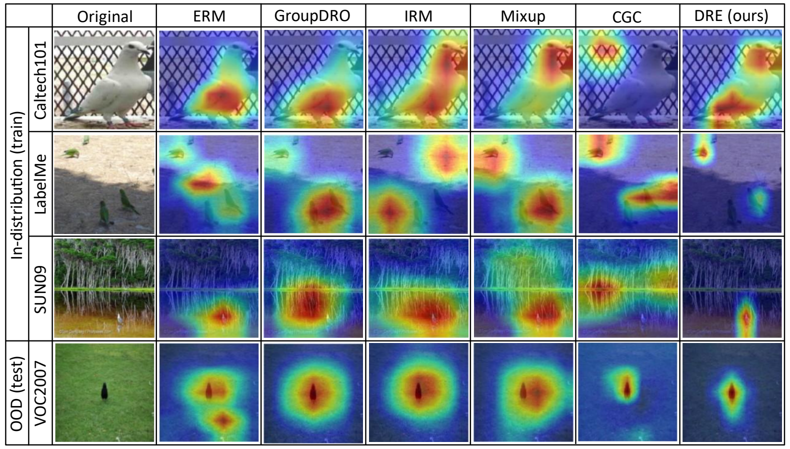
Unreliable explanations on OOD data.



Explanation qualities gap between ID and OOD data.

Observation: Data-driven explanations are ***NOT*** robust against OOD data.

RQ3. *Can robust explanations benefit the model's generalization capability?*



Our model alleviates the reliance on spurious correlations.

- Explanation
  - ↑ 4.7% (Fidelity)
  - ↑ 29.2% (Scientific Consist.)
  - ↑ 16.6% (Distributional Consist.)
- Prediction
  - ↑ 5.1% (Terra Incognita)
  - ↑ 1.0% (VLCS)
  - ↑ 18.5% (Urban Land)

Explainability Demand for ML Models

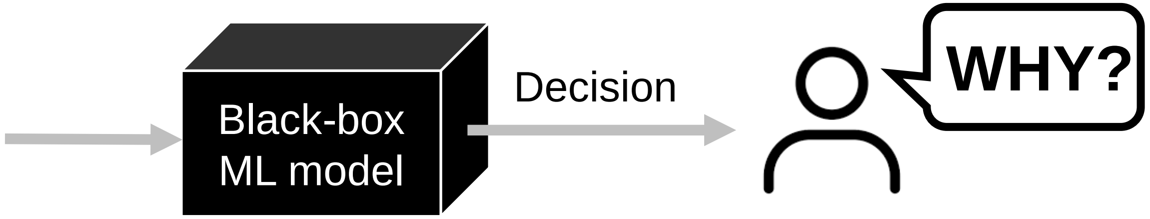

Healthcare


Criminal Justice

Finance


Self-driving Cars

...



The absence of reliable explanations can lead to ***severe consequences***.

- Catastrophic outcomes in ***high-stakes*** applications.


Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017


ARTIFICIAL INTELLIGENCE | OPINION

Who Is Liable When AI Kills?

We need to change rules and institutions while still promoting innovation to protect people from faulty AI

By George Maliha, Ravi B. Parikh on June 29, 2022


COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

- Violation of ***regulations***.

GDPR.EU

This project is co-funded by the Horizon 2020 Framework Programme of the European Union

"*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*"

Article.22 of General Data Protection Regulation (GDPR) empowers individual with **the right to demand explanation of an AI system**.

[Lakkaraju et al. 2023]

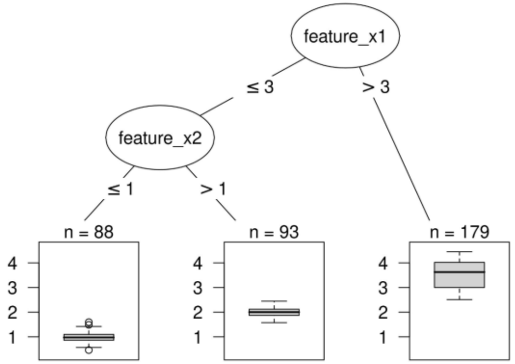
T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

Related Work: Explainable Machine Learning (XML)

<u>Intrinsic XML methods</u>: Explanations are inherent to the model architecture and training.

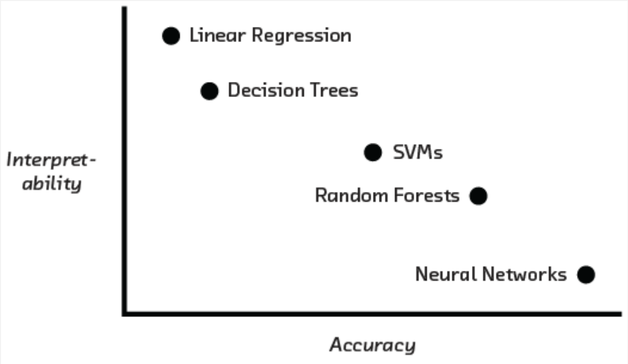
- Linear regression
- Logistic regression
- Decision trees
- RuleFit
- Naive Bayes
- K-nearest neighbors

…



Decision Tree

Limitation: *<u>Interpretability-accuracy Trade-off</u>*

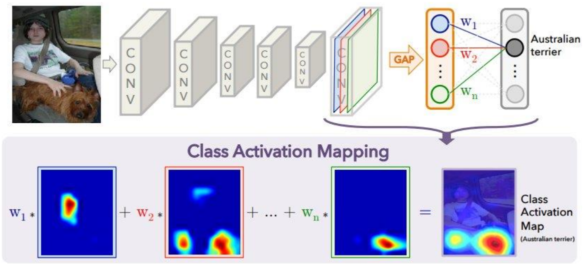


[Cireşan et al. 2012, Caruana et al. 2006, Frosst et al. 2017, Stewart et al. 2020]

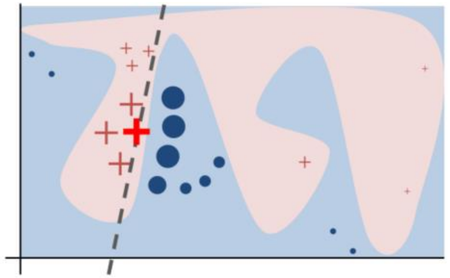
<u>Post-hoc XML methods</u>: Provide explanations for a pre-built model in a post-hoc manner.

- Occlusion Sensitivity [Zeiler et al. 2014]
- Class Activation Map (CAM) [Zhou et al. 2016]
- Layer-Wise Relevance Propagation (LRP) [Bach et al. 2015]
- Integrated Gradients (IG) [Sundararajan et al. 2016]
- Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro et al. 2016]
- Shapley Additive Explanations (SHAP) [Lundberg et al. 2017]

…



CAM



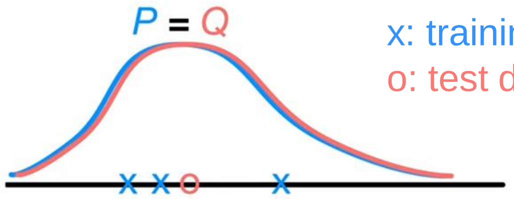
LIME

Limitation: Post-hoc explanations are *<u>not robust</u>* against distributional shifts, *<u>unreliable</u>* on ***out-of-distribution*** data.

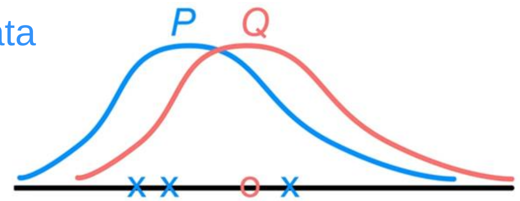
T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

Out-of-distribution (OOD) Challenges

- A highly accurate model on average can ***fail catastrophically*** on OOD data.



x: training data
o: test data

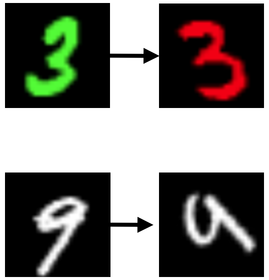


Expectation: Same distribution (i.i.d.)

Reality: Distributional drifts

Spurious correlation

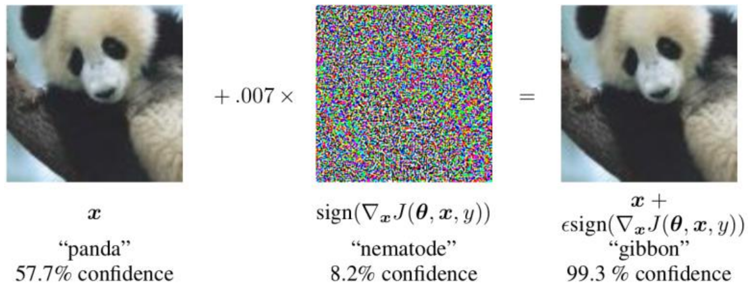
Color and rotation are irrelevant features to digits.



[Gulrajani'21]

Adversarial attacks

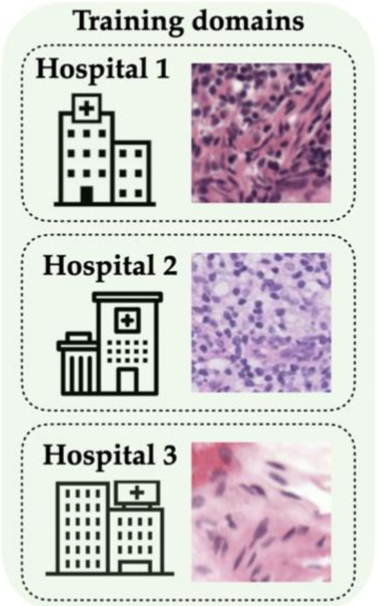
By adding imperceptibly small noise, classification results can be significantly changed.



[Goodfellow'15]

Sub-populations

Cell images are distributed differently in different hospitals.



[Robey'21]

Naturally-occurring variation

Distribution shift caused by seasons, weather, and geographical locations.

Sunny

Rainy

Fog

Snow

The robustness of ***explanations*** against OOD data remains a vital yet seldom-investigated question.

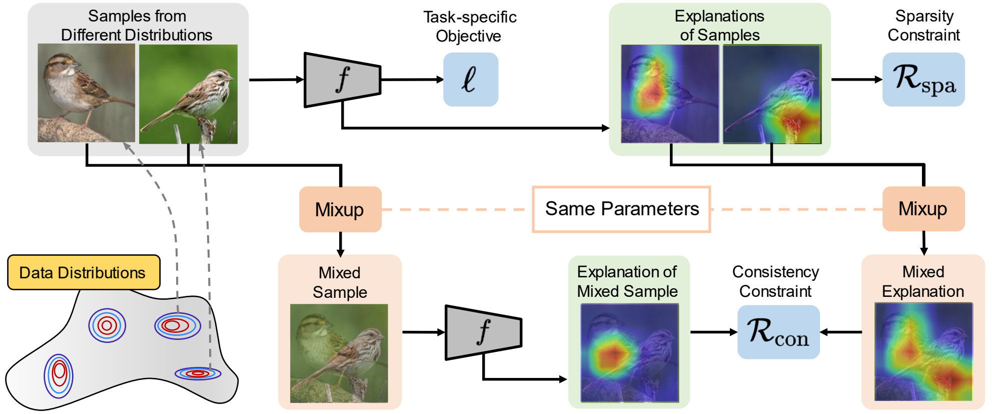
T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

Outline

Highlights

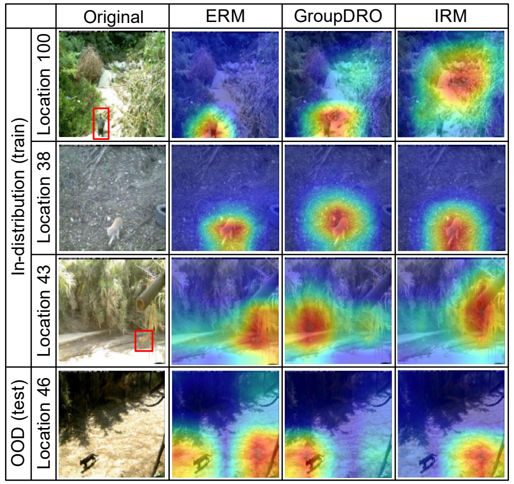
- *Observation*: data-driven explanations are ***unreliable*** on out-of-distribution (OOD) data.
- *Method*: framework **Distributionally Robust Explanations (DRE)** for the learning of consistent explanations across distributions.
- *Experiments*: when testing on OOD data, our model significantly improve the explanation fidelity by ***6.0%*** and prediction accuracy by ***6.9%*** on Terra Incognita dataset.
- Code and pre-trained weights: https://github.com/tangli-udel/DRE

RQ2. *How to develop robust explanations against out-of-distribution data?*

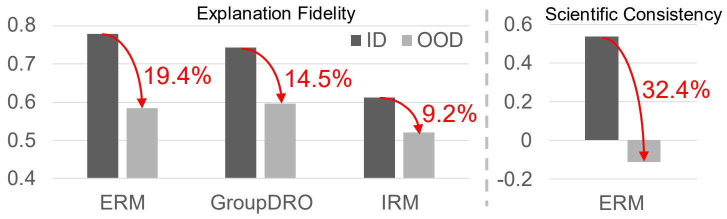


Overview of the proposed Distributionally Robust Explanation (DRE).

RQ1. *Are data-driven explanations robust against out-of-distribution data?*



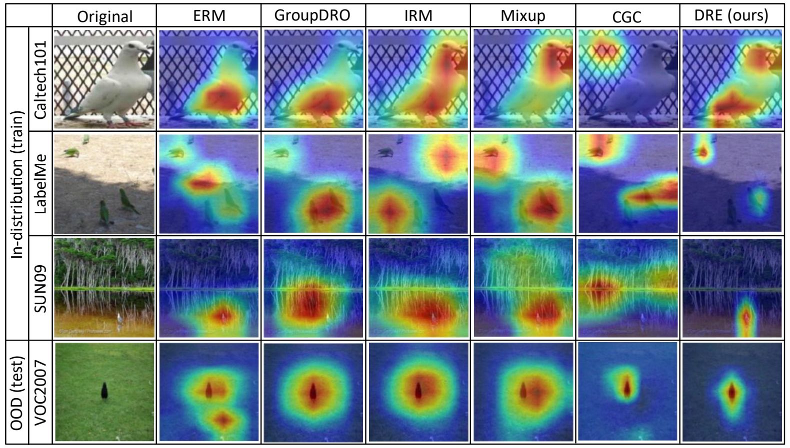
Unreliable explanations on OOD data.



Explanation qualities gap between ID and OOD data.

Observation: Data-driven explanations are ***NOT*** robust against OOD data.

RQ3. *Can robust explanations benefit the model's generalization capability?*



Our model alleviates the reliance on spurious correlations.

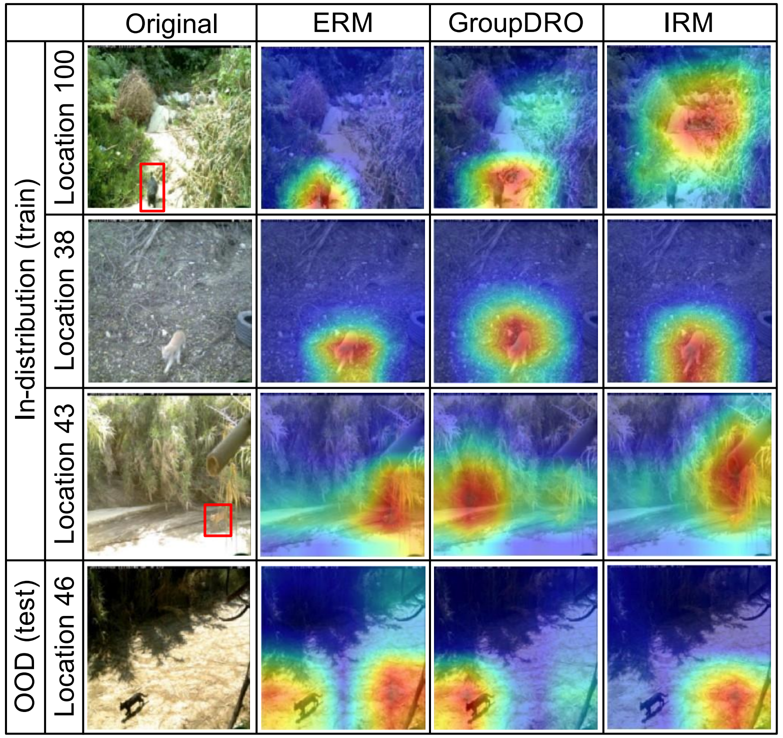
- Explanation
  - ↑ 4.7% (Fidelity)
  - ↑ 29.2% (Scientific Consist.)
  - ↑ 16.6% (Distributional Consist.)
- Prediction
  - ↑ 5.1% (Terra Incognita)
  - ↑ 1.0% (VLCS)
  - ↑ 18.5% (Urban Land)

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ1. Are data-driven explanations robust against out-of-distribution data?

--- Observations

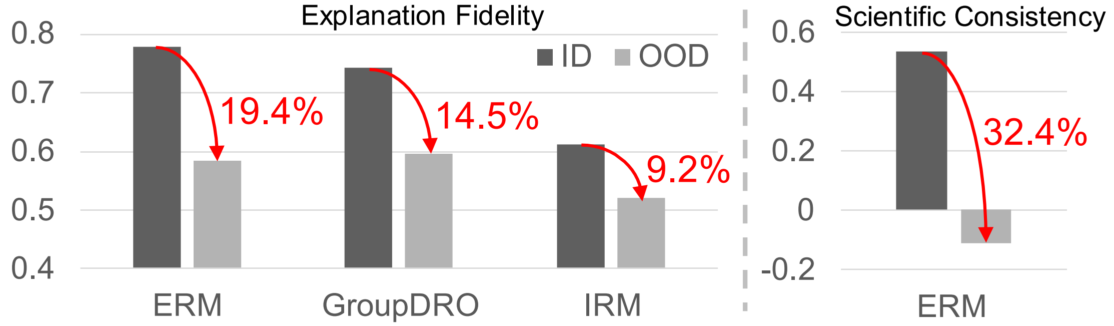
Qualitatively:



Grad-CAM visualization of images from different distributions in Terra Incognita dataset. Models are trained using representative OOD generalization methods.

- Even with correct predictions, the explanations would also highlight ***background pixels*** (e.g., tree branches) on OOD data.

Quantitatively:



Fidelity evaluation on Grad-CAM explanations of images from Terra Incognita dataset; scientific consistency evaluation on Input Gradient explanations of tabular data from Urban Land dataset.

- The explanation quality experiences a ***severe drop*** on OOD data, in terms of fidelity and scientific consistency.

Takeaway 1:

- Data-driven explanations are ***NOT*** robust against OOD data.
- The explanations excessively relied on ***spurious correlations***.

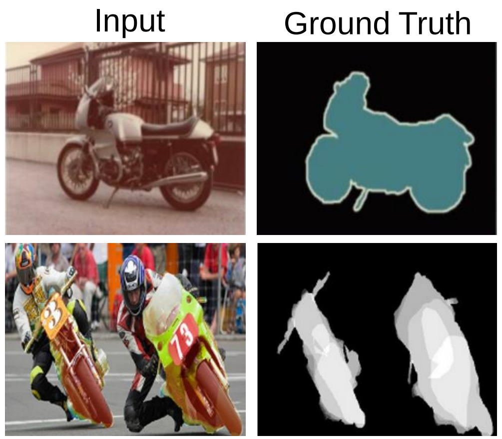
T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ2. How to develop robust explanations against out-of-distribution data?

--- The Gap of Supervision

In order to alleviate the reliance on spurious correlations, ***supervision of explanations*** are essential.
They are typically derived from:

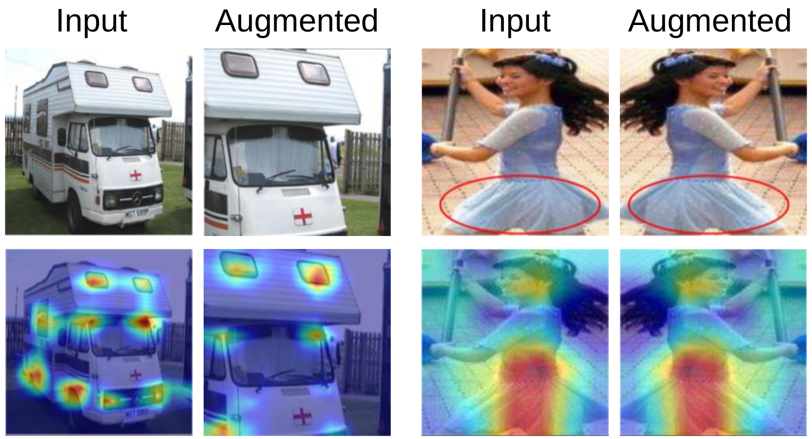
Ground Truth



Explanation annotations.

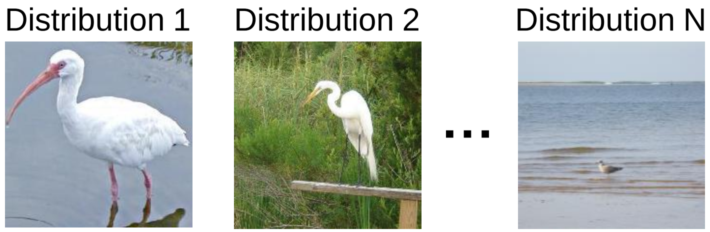
[Selvaraju et al. 2017, Mohseni et al. 2021]

Data Augmentation



One-to-one mapping between image transforms.

[Guo et al. 2019, Pillai et al. 2022]



Real-world Distributional Shifts.

<u>Supervision Gaps</u>:

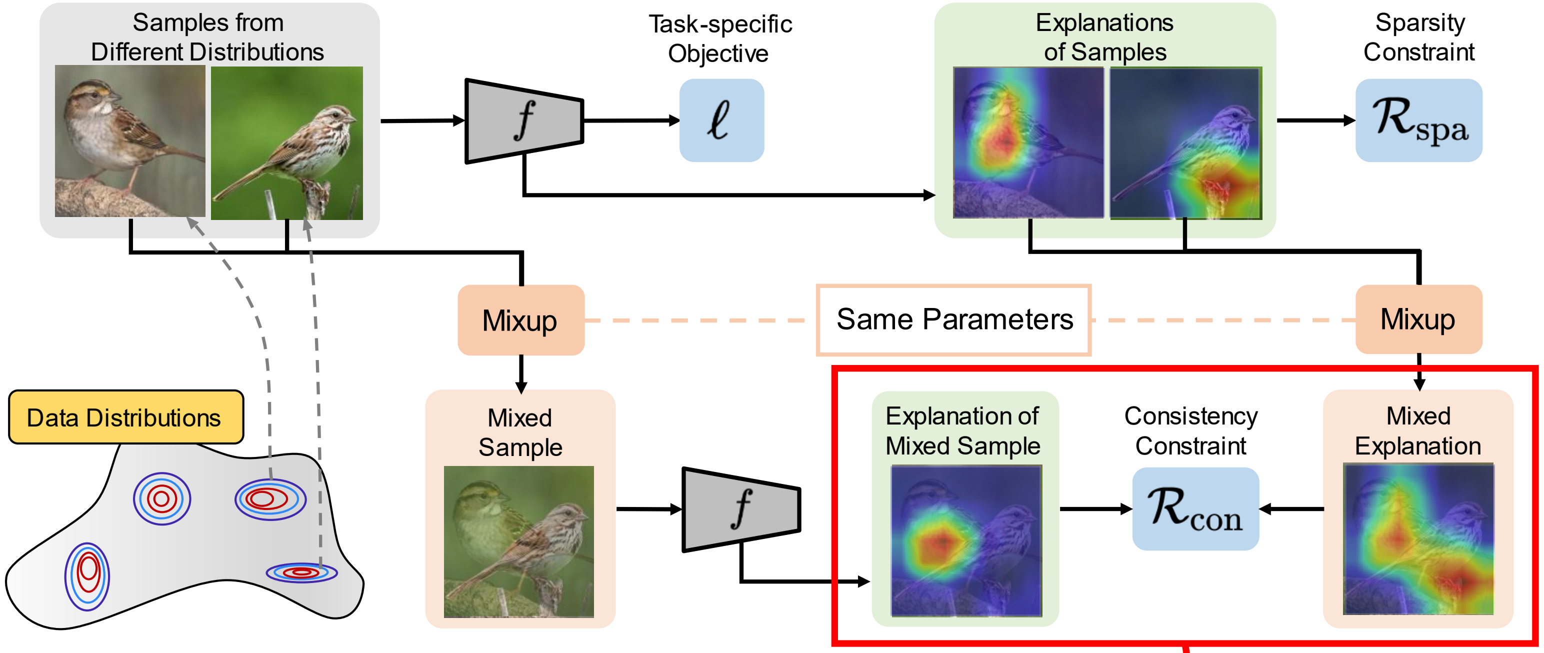
Obtaining ground truth explanation annotations are ***prohibitively expensive*** or even ***impossible***.

Naturally-occurring distributional shifts are different from data augmentations, there is ***no one-to-one mapping*** between explanations.

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ2. How to develop robust explanations against out-of-distribution data?

--- Our Solution: Distributionally Robust Explanations (DRE)



Key idea: leveraging the ***mixed explanation*** between distributions to provide supervisory signals for the learning of explanations.

Merits:

- Providing supervisory signals for the learning of explanations ***without human annotation***.
- Achieving a simple but effective ***inter-distributional transformation***.

$$\min_{f\in\mathcal{F}} \mathcal{R}(f) \quad \mathrm{s.t.}\ \mathcal{D}[g(\mathcal{M}(\mathbf{x}_e,\mathbf{x}_{e'})), \mathcal{M}(g(\mathbf{x}_e), g(\mathbf{x}_{e'}))] \leq \epsilon$$

Explanation of mixed sample → consistent ← Mixed explanations

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ3. Can robust explanations benefit the model's generalization capability?

--- Experiments: Image Classification

Data and Distributions:

✓ **Terra Incognita** ([Beery et al 2019]) In the wild camera trap images.

Distributions: Camera Locations with different illumination, perspective, etc.

Location 100 Location 38 Location 43 Location 46

✓ **VLCS** ([Fang et al. 2013]) Natural images from different sub-datasets.

Distributions: Sub-datasets with different styles, backgrounds, etc.

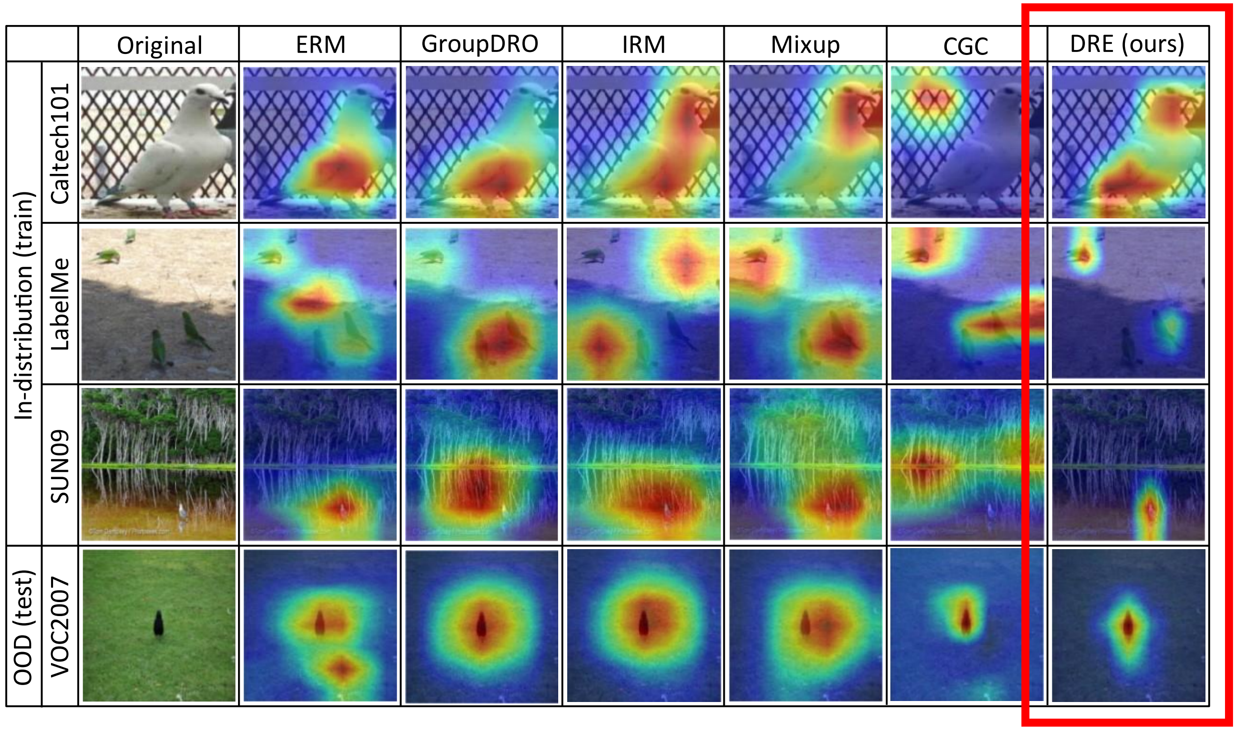
Caltech101 LabelMe SUN09 VOC2007

Metrics:

✓ **Distributional Consistency:** Measuring the explanation consistency between *in-* and *out-of-distribution* data.

✓ **Explanation Fidelity** ([Petsiuk et al. 2018])**:** Measuring how well an explanation reflects underlying decision-making process.

Qualitative Results on VLCS:



➢ Our method alleviates the model's reliance on background pixels and ensures ***consistent explanations*** across distributions.

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ3. Can robust explanations benefit the model’s generalization capability?

--- Experiments: Image Classification

Data and Distributions:

- ✓ **Terra Incognita** ([Beery et al 2019]) In the wild camera trap images.
  Distributions: Camera Locations with different illumination, perspective, etc.


Location 100


Location 38


Location 43


Location 46

- ✓ **VLCS** ([Fang et al. 2013]) Natural images from different sub-datasets.
  Distributions: Sub-datasets with different styles, backgrounds, etc.


Caltech101


LabelMe


SUN09


VOC2007

Metrics:

- ✓ **Distributional Consistency:** Measuring the explanation consistency between *in-* and *out-of-distribution* data.
- ✓ **Explanation Fidelity** ([Petsiuk et al. 2018])**:** Measuring how well an explanation reflects underlying decision-making process.

Improvements of DRE on OOD Explanations:

Consistency	↑ to ERM	↑ to GroupDRO	↑ to IRM	↑ to CGC
Terra	77.9%	77.1%	74.6%	16.6%
VLCS	71.4%	67.0%	63.2%	69.9%

Fidelity	↑ to ERM	↑ to GroupDRO	↑ to IRM	↑ to CGC
Terra	6.0%	4.7%	12.4%	34.6%
VLCS	2.1%	7.4%	8.0%	20.8%

Improvements of DRE on OOD Accuracy:

Accuracy	↑ to ERM	↑ to GroupDRO	↑ to IRM	↑ to CGC
Terra	6.9%	9.8%	5.4%	5.1%
VLCS	2.0%	2.8%	1.0%	3.4%

Takeaway 2:

- Robust explanations ***significantly benefit*** the model’s generalization capability by alleviating its reliance on ***spurious correlations***.

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ3. Can robust explanations benefit the model's generalization capability?

--- Ablation Study

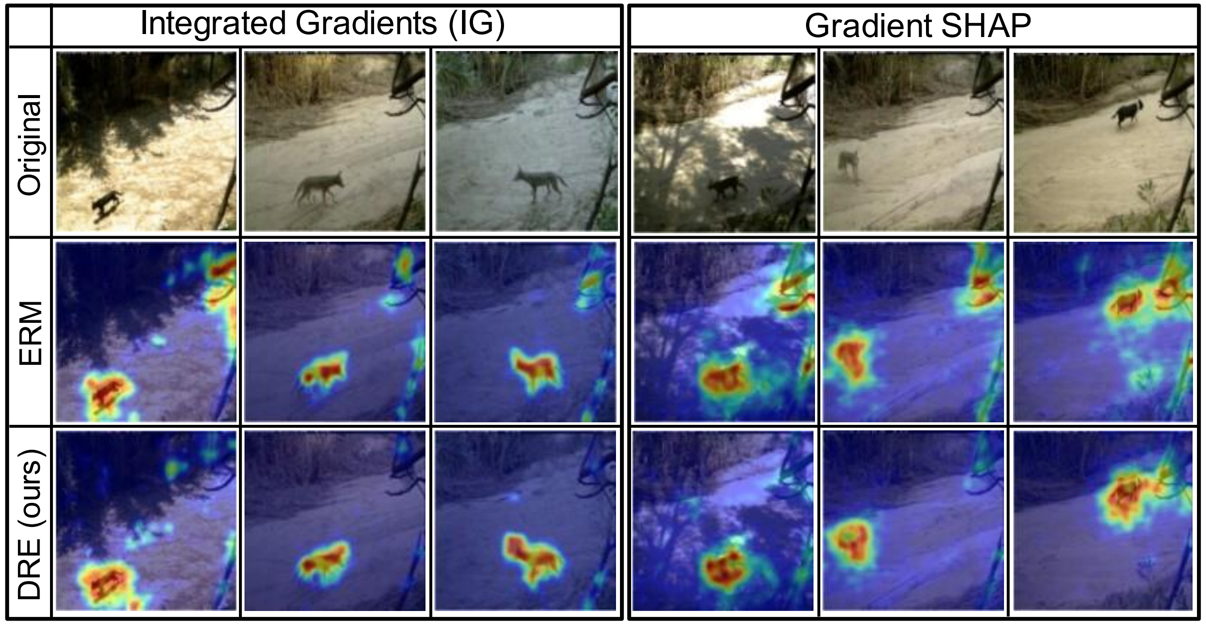
Ablation Study on VLCS:

↑ to ERM	Consistency	Fidelity	Accuracy
DRE w/o reg.	9.1%	-0.2%	3.5%
DRE w/o consist.	63.6%	-6.8%	-3.4%
DRE (full)	26.7%	1.4%	5.6%

Blindly imposing constraints on consistency or sparsity would deteriorate accuracy or explanation quality on OOD data.

- ➢ Our method strikes a ***good balance*** between optimization objectives.

IG and SHAP Visualizations on Terra:



The saliency maps of our model alleviate the reliance on background pixels, and clearly depicts the contour of the object on OOD data.

- ➢ The advanced explainability of our model ***can be generalized*** to other data-driven explanation methods.

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

RQ3. Can robust explanations benefit the model's generalization capability?

--- Experiments: Regression on Scientific Tabular Data

Data and Distributions:

✓ **Urban Land** ([Gao et al 2019]) A large-scale spatiotemporal dataset used for urban land fraction prediction.

Distributions: continental regions with different topographic, population, and historical urban fraction conditions.

Metrics:

✓ **Distributional Consistency:** Measuring the explanation consistency between *in-* and *out-of-distribution* data.

✓ **Scientific Consistency:** The consistency between explanations and ground truth domain knowledge or principles.

Improvements of DRE on OOD Explanations:

↑ to ERM	DRE (ours)
Explanation consist.	84.5%
Scientific consist.	**29.2%**

Improvements of DRE on OOD Accuracy:

↑ to ERM	DRE (ours)
Regression accuracy	18.5%

The results are on the average of different experiment trials that hold out each distribution as the OOD testing set.

Takeaway 3:

- Robust explanations significantly improves the ***scientific consistency*** on OOD data.

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

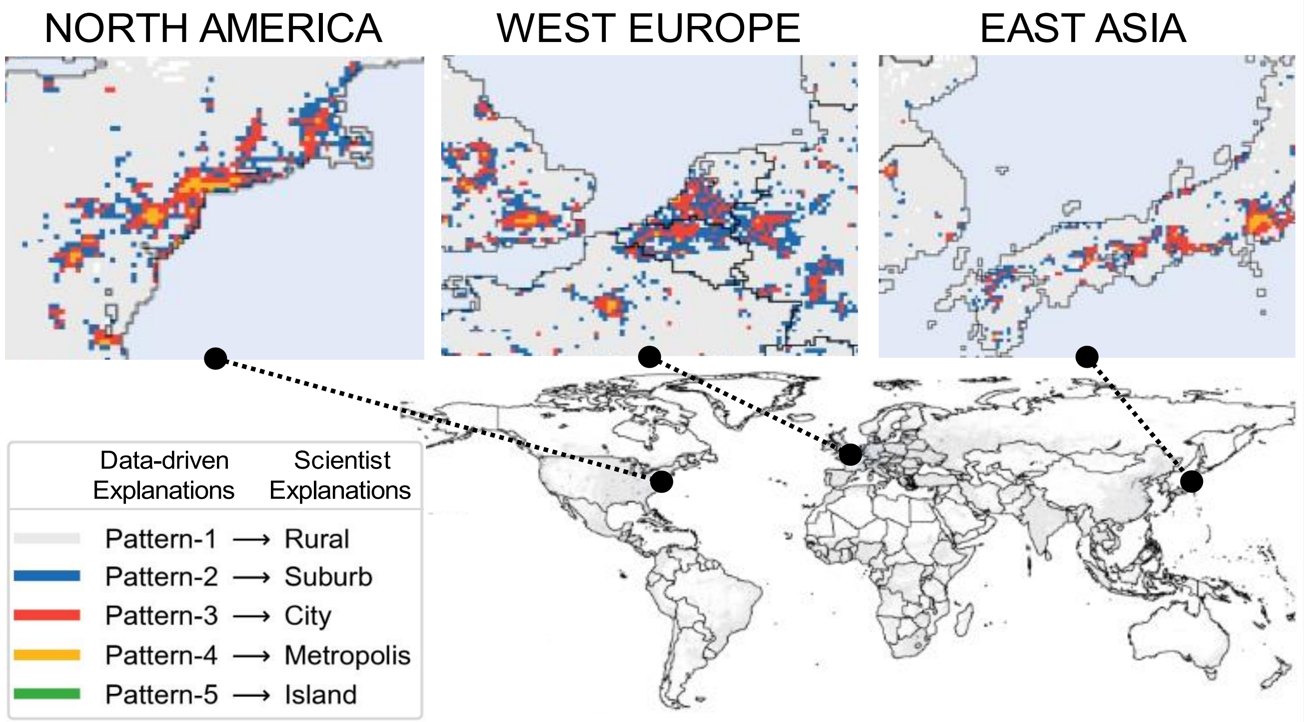
Broader Impact

--- Can robust explanations catalyze scientific knowledge discovery?

<u>Scientific Knowledge Discovery via DRE:</u>

- Explanations on unseen distributions can reveal ***unknown patterns*** arising from local data.
- The ***discrepancy*** between consistent (DRE) and inconsistent explanations can be exploited for knowledge discovery.

Our preliminary work on this topic won the ***Best Paper Award*** in the *Machine Learning in Public Health* workshop at NeurIPS 2021.



T. Li, J. Gao, and X. Peng. *Deep learning for spatiotemporal modeling of urbanization*. NeurIPS'21W

An instance of recently uncovered domain knowledge: five distinct patterns of urbanization. A human-in-the-loop explanation system that integrates domain experts and data-driven explanations can be developed to detect and analyze the explanation discrepancy.

T. Li, F. Qiao, M. Ma, and X. Peng. *Are Data-driven Explanations Robust against Out-of-distribution Data?* CVPR 2023.

Thank you!

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

Tang Li   Fengchun Qiao   Mengmeng Ma   Xi Peng
Department of Computer & Information Sciences
University of Delaware

Poster: TUE-AM-364