

# Supplementary Materials: Causality Inspired Representation Learning for Domain Generalization

Fangrui Lv<sup>1</sup> Jian Liang<sup>2</sup> Shuang Li<sup>1,\*</sup> Bin Zang<sup>1</sup> Chi Harold Liu<sup>1</sup> Ziteng Wang<sup>3</sup> Di Liu<sup>2</sup>

<sup>1</sup> Beijing Institute of Technology, China <sup>2</sup> Alibaba Group, China <sup>3</sup> Yizhun Medical AI Co., Ltd, China

<sup>1</sup> {fangruilv, shuangli, binzang}@bit.edu.cn, liuchi02@gmail.com  
<sup>2</sup> {xuelang.lj, wendi.ld}@alibaba-inc.com <sup>3</sup> ziteng.wang@yizhun-ai.com

## 1. Potential Negative Societal Impacts

Our work focuses on domain generalization and attempts to excavate the intrinsic causal mechanisms from a causal view, which further enhances the generalization capability of the learned model on OOD distribution. This approach exerts a positive influence on the society and the community for saving the cost and time of data annotation, boosting the reusability of knowledge across domains, and greatly improving the generalization ability. However, this work may also suffer from some negative impacts, which is worthy of further research and exploration. Specifically, more jobs of classification or target detection for conditions that out of the support of observed data distributions may be cancelled. What's more, we need to be cautious about the reliability of the system, which might be misleading when using in some conditions that are very far from the observed distributions.

## 2. Implementation Details

### 2.1. Z-score Normalization

As mentioned in the section *Causal Intervention Module*, before measuring the correlation of representations before and after the intervention upon non-causal factors, we conduct Z-score normalization on the columns of representations  $\mathbf{R}^o \in \mathbb{R}^{B \times N}$  and  $\mathbf{R}^a$  as follows, which can convert data of different orders of magnitude into unified Z-score measurement for fair comparison:

$$\begin{aligned}\tilde{\mathbf{r}}_i^o &= \frac{\mathbf{r}_i^o - \frac{1}{N} \sum_{i=1}^N \mathbf{r}_i^o}{\sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{r}_i^o - \frac{1}{N} \sum_{i=1}^N \mathbf{r}_i^o)^2}}, \\ \tilde{\mathbf{r}}_i^a &= \frac{\mathbf{r}_i^a - \frac{1}{N} \sum_{i=1}^N \mathbf{r}_i^a}{\sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{r}_i^a - \frac{1}{N} \sum_{i=1}^N \mathbf{r}_i^a)^2}},\end{aligned}\quad (1)$$

\* Corresponding author.

where  $\tilde{\mathbf{r}}_i^o, \tilde{\mathbf{r}}_i^a$  denote the  $i$ -th column of  $\mathbf{R}^o$  and  $\mathbf{R}^a$  respectively.

### 2.2. Gumbel-Softmax Trick

We apply the commonly-used Gumbel-softmax [9] trick to approximately sample a  $k$ -hot vector, where  $k = \lfloor \kappa N \rfloor \in \mathbb{Z}_+$ . Specifically, in Eq. (12), let  $\mathbf{z} = \hat{\mathbf{w}}(\mathbf{r}) \in \mathbb{R}^N$  be a probability vector, where for  $j \in \{1, \dots, N\}$ ,  $z_j \geq 0$  and  $\sum_j z_j = 1$ . Then we define the sampled vector  $\mathbf{m} = \text{Gumbel-Softmax}(\hat{\mathbf{w}}(\mathbf{r}), \kappa N) \in \mathbb{R}^N$ , where for a pre-defined  $\tau > 0, j \in \{1, \dots, N\}, l \in \{1, \dots, k\}$ ,

$$\begin{aligned}m_j &= \max_{l \in \{1, \dots, k\}} \frac{\exp((\log z_j + \xi_j^l)/\tau)}{\sum_{j'=1}^N \exp((\log z_{j'} + \xi_{j'}^l)/\tau)}, \\ \xi_j^l &= -\log(-\log u_j^l), u_j^l \sim \text{Uniform}(0,1).\end{aligned}\quad (2)$$

We follow a common setting [4] to let  $\tau = 0.5$ .

### 2.3. Network Structure

For Digits-DG, the network is the same as [25] which constructs the network with four  $3 \times 3$  convolutional layers (each followed by ReLU and  $2 \times 2$  max-pooling) and a softmax classification layer, denoted as ConvNet. As for PACS and Office-Home, we adopt ResNet [7] pre-trained on ImageNet [5] as the backbone. And the masker is implemented by a 3-layer MLP which is randomly initialized. For fairness, the feature dimension  $N$  is 256, 512, and 2048 for the experiments with backbone ConvNet, ResNet-18, and ResNet-50 respectively following previous works.

## 3. Additional Results

### 3.1. Sensitivity to Feature Dimension

To analyze the influence of feature dimension, we first conduct experiments for different feature dimensions  $N = \{128, 256, 512, 1024, 2048\}$  with a fixed  $\kappa = 60\%$  on PACS

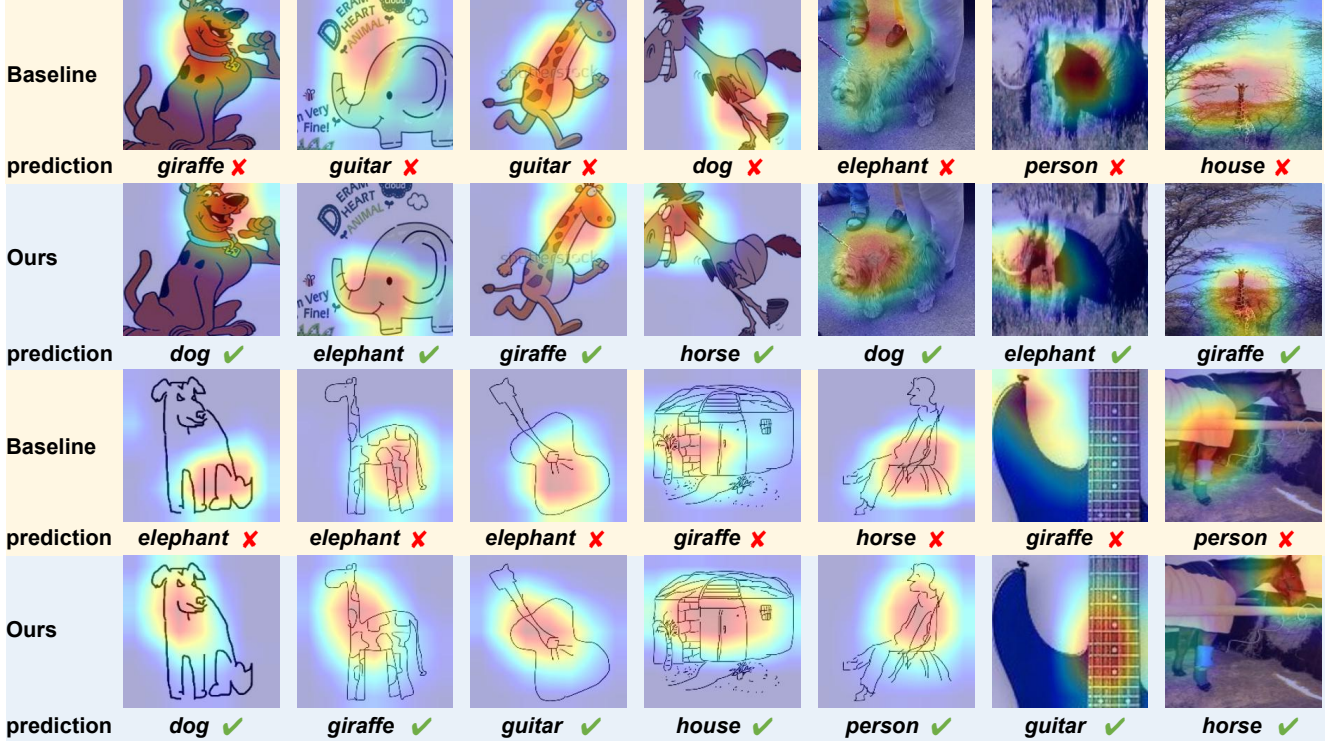


Figure 1. More examples of the visualization explanation on PACS dataset, with Cartoon, Photo, Sketch as the unseen target domain respectively for baseline (i.e., DeepAll) and CIRL methods.

Table 1. Model performance sensitivity to feature dimension (ResNet-18).

$N$	Art	Cartoon	Photo	Sketch	Avg.
128	85.9	80.1	95.9	82.6	86.12
256	85.8	80.2	<b>96.4</b>	82.6	86.25
512	<b>86.1</b>	<b>80.6</b>	95.9	<b>82.7</b>	<b>86.32</b>
1024	85.9	80.4	95.7	<b>82.7</b>	86.17
2048	86.0	80.2	96.3	82.4	86.23

to explore the influence of feature dimension on model performance, as Table 1 shows. It is clear that the performance of model remains relatively stable when the feature dimension varies, which demonstrates the stability of our method. And then we conduct experiments for fixed feature dimensions  $N = 512$  and  $N = 2048$  with different  $\kappa = \{50\%, 60\%, 70\%, 80\%, 90\%\}$  on PACS to investigate the influence of feature dimension on the choice of  $\kappa$ , the results are shown in Table 2. We can see that  $\kappa$  is shown not sensitive to feature dimension,  $\kappa = 60\%$  is a stable choice for different feature dimensions.

### 3.2. Sampling Strategies for Amplitude Mixing

Table 2.  $\kappa$  selection sensitivity to feature dimension (ResNet-18).

$\kappa$	Art	Cartoon	Photo	Sketch	Avg.
$N = 512$					
50%	85.5	79.5	<b>96.2</b>	82.0	85.80
60%	<b>86.1</b>	80.6	95.9	<b>82.7</b>	<b>86.32</b>
70%	85.7	<b>80.8</b>	95.8	82.1	86.13
80%	85.4	79.6	95.6	81.8	85.60
90%	85.3	79.6	95.6	81.4	85.48
$N = 2048$					
50%	85.7	79.8	96.2	<b>82.6</b>	86.08
60%	<b>86.0</b>	<b>80.2</b>	<b>96.3</b>	82.4	<b>86.23</b>
70%	85.6	80.1	96.0	81.7	85.85
80%	84.7	78.9	95.9	82.1	85.40
90%	84.6	79.7	95.8	81.3	85.35

In the causal intervention module, we mix amplitude spectrum of a specific image and another image sampled randomly from arbitrary source domains. Nevertheless, we can also restrict the sample pair to be taken from the same domain (intra-domain) or different domains (inter-domain). To explore the effect of different sampling strategies, we

Table 3. Impact of sampling strategies for amplitude mixing.

Sampling Strategy	Art	Cartoon	Photo	Sketch	Avg.
intra-domain	84.92	80.76	96.23	82.77	86.17
inter-domain	85.45	79.93	96.53	82.14	86.01
random	86.08	80.59	95.93	82.67	86.32

Table 4. Leave-one-domain-out results on mini-DomainNet with ResNet-18.

Method	Clipart	Painting	Real	Sketch	Avg.
ERM [21]	65.5	57.1	62.3	57.1	60.50
DRO [18]	64.8	57.4	61.5	56.9	60.15
Mixup [24]	67.1	59.1	64.3	59.2	62.42
MLDG [11]	65.7	57.0	63.7	58.1	61.12
CORAL [16]	66.5	59.5	66.0	59.5	62.87
MMD [10]	65.0	58.0	63.8	58.4	61.30
MTL [2]	65.3	59.0	65.6	58.5	62.10
SagNet [15]	65.0	58.1	64.2	58.1	61.35
CIRL (ours)	<b>70.2</b>	<b>62.9</b>	<b>67.8</b>	<b>63.6</b>	<b>66.13</b>

conduct experiments by using only intra-domain or inter-domain strategy on PACS with ResNet-18, as Table 3 shows. As we can see, CIRL is shown not sensitive to the sampling strategies. Whether intra- or inter-domain sampling strategy brings a good performance, and using a fully random strategy works best, perhaps because more causal interventions are conducted which leads to extracting better causal features.

### 3.3. Experimental Results on mini-DomainNet

Note that the benchmarks used for experiments in the main body are all relatively small-scale datasets, which may be fairly saturated in performance so that the improvement of our CIRL seems to be incremental. Thus, we additionally carry out experiments on a large-scale dataset mini-DomainNet following [27] which has 140K images with 126 classes. The results are shown in Table 4, we can see that CIRL outperforms the SOTA method by a large margin of 3.26%, which validates our superiority.

### 3.4. More Examples for Visual Explanation

We present more examples of attention maps of the last convolutional layer for baseline (i.e., DeepAll) and CIRL methods in Fig. 1, utilizing the visualization technique in [19]. As we can see, in all the tasks, the representations learned by CIRL precisely capture the category-related information of different objects that contributes to the classification, while the representations learned by baseline method contain many noisy information such as background that leads to wrong predictions.

### 3.5. Experimental Results with Error Bars

For the sake of objective, we run all the experiments multiple times with random seed. We report the average

results in the main body of paper for elegant, and show the complete results with error bars in the form of  $\text{mean} \pm \text{std}$  below (Table. 5,6,7,8).

## References

- [1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, pages 1006–1016, 2018. 4
- [2] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 22:2:1–2:55, 2021. 3
- [3] Fabio Maria Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019. 4, 5
- [4] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892, 2018. 1
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1
- [6] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6447–6458, 2019. 4
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1
- [8] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140, 2020. 4, 5
- [9] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 1
- [10] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, pages 10285–10295, 2019. 3
- [11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, pages 3490–3497, 2018. 3
- [12] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 4, 5

Table 5. Leave-one-domain-out results on Digits-DG.

Methods	MNIST	MNIST-M	SVHN	SYN	Avg.
Jigen [3]	96.5	61.4	63.7	74.0	73.9
L2A-OT [26]	96.7	63.9	68.6	83.2	78.1
DeepAll [25]	95.8±0.3	58.8±0.5	61.7±0.5	78.6±0.6	73.7
CCSA [14]	95.2±0.2	58.2±0.6	65.5±0.2	79.1±0.8	74.5
MMD-AAE [12]	96.5±0.1	58.4±0.1	65.0±0.1	78.4±0.2	74.6
CrossGrad [20]	96.7±0.1	61.1±0.5	65.3±0.5	80.2±0.2	75.8
DDAIG [25]	96.6±0.2	64.1±0.4	68.6±0.6	81.0±0.5	77.6
FACT [23]	<b>97.9±0.2</b>	65.6±0.4	72.4±0.7	<b>90.3±0.1</b>	81.5
CIRL ( <i>ours</i> )	96.08±0.2	<b>69.87 ± 0.5</b>	<b>76.17 ± 0.4</b>	87.68±0.3	<b>82.5</b>

Table 6. Leave-one-domain-out results on PACS with ResNet-18.

Methods	Art	Cartoon	Photo	Sketch	Avg.
JiGen [3]	79.42	75.25	96.03	71.35	80.51
L2A-OT [26]	83.30	78.20	96.20	73.60	82.80
RSC [8]	83.43	80.31	95.99	80.85	85.15
DeepAll [25]	77.63±0.84	76.77±0.33	95.85±0.20	69.50±1.26	79.94
MetaReg [1]	83.70±0.19	77.20±0.31	95.50±0.24	70.30±0.28	81.70
DDAIG [25]	84.20±0.30	78.10±0.60	95.30±0.40	74.70±0.80	83.10
CSD [17]	78.90±1.10	75.80±1.00	94.10±0.20	76.70±1.20	81.40
MASF [6]	80.29±0.18	77.17±0.08	94.99±0.09	71.69±0.22	81.04
EISNet [22]	81.89±0.88	76.44±0.31	95.93±0.06	74.33±1.37	82.15
MatchDG [13]	81.32±0.38	<b>80.70 ± 0.54</b>	96.53±0.05	79.72 ± 1.01	84.56
FACT [23]	85.90±0.27	79.35±0.03	<b>96.61±0.17</b>	80.89±0.26	85.69
CIRL ( <i>ours</i> )	<b>86.08±0.32</b>	80.59±0.19	95.93±0.03	<b>82.67±0.47</b>	<b>86.32</b>

Table 7. Leave-one-domain-out results on PACS with ResNet-50.

Methods	Art	Cartoon	Photo	Sketch	Avg.
RSC [8]	87.89	82.16	97.92	83.35	87.83
DeepAll [25]	84.94±0.66	76.98±1.13	97.64±0.10	76.75±0.41	84.08
MetaReg [1]	87.20±0.13	79.20±0.27	97.60±0.31	70.30±0.18	83.60
MASF [6]	82.89±0.16	80.49±0.21	95.01±0.10	72.29±0.15	82.67
EISNet [22]	86.64±1.41	81.53±0.64	97.11±0.40	78.07±1.43	85.84
MatchDG [13]	85.61±0.81	82.12±0.69	<b>97.94±0.27</b>	78.76±1.13	86.11
FACT [23]	<b>90.89±0.19</b>	83.65±0.12	97.78±0.05	86.17±0.14	89.62
CIRL ( <i>ours</i> )	90.67±0.21	<b>84.30±0.17</b>	97.84±0.08	<b>87.68±0.40</b>	<b>90.12</b>

[13] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pages 7313–7324, 2021. 4

[14] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5716–5726, 2017. 4, 5

[15] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *CoRR*, abs/1910.11645, 2019. 3

[16] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–

Table 8. Leave-one-domain-out results on Office-Home with ResNet-18.

Methods	Art	Clipart	Product	Real	Avg.
Jigen [3]	53.04	47.51	71.47	72.79	61.20
RSC [8]	58.42	47.90	71.63	74.54	63.12
L2A-OT [26]	60.60	50.10	74.80	<b>77.00</b>	65.60
DeepAll [25]	57.88±0.20	52.72±0.50	73.50±0.30	74.80±0.10	64.72
CCSA [14]	59.90±0.30	49.90±0.40	74.10±0.20	75.70±0.20	64.90
MMD-AAE [12]	56.50±0.40	47.30±0.30	72.10±0.30	74.80±0.20	62.70
CrossGrad [20]	58.40±0.70	49.40±0.40	73.90±0.20	75.80±0.10	64.40
DDAIG [25]	59.20±0.10	52.30±0.30	74.60±0.30	76.00±0.10	65.50
FACT [23]	60.34±0.11	54.85±0.37	74.48±0.13	76.55±0.10	66.56
CIRL ( <i>ours</i> )	<b>61.48±0.17</b>	<b>55.28±0.29</b>	<b>75.06±0.24</b>	76.64±0.09	<b>67.12</b>

- 1415, 2019. 3
- [17] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, pages 7728–7738, 2020. 4
- [18] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019. 3
- [19] 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 *ICCV*, pages 618–626, 2017. 3
- [20] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018. 4, 5
- [21] Vladimir Vapnik. An overview of statistical learning theory. *IEEE Trans. Neural Networks*, 10(5):988–999, 1999. 3
- [22] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, volume 12354, pages 159–176, 2020. 4
- [23] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021. 4, 5
- [24] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3
- [25] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020. 1, 4, 5
- [26] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578, 2020. 4, 5
- [27] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Trans. Image Process.*, 30:8008–8018, 2021. 3