Interpret Your Decision: Logical Reasoning Regularization for Generalization in Visual Classification (NeurIPS24 Spotlight)

Zhaorui Tan [1,2], Xi Yang [1,*], Qiufeng Wang[1], Anh Nguyen[2], Kaizhu Huang [3,*]

[1]Department of Intelligent Science, Xi'an Jiaotong-Liverpool University
[2]Department of Computer Science, University of Liverpool
[3] Data Science Research Center, Duke Kunshan University,
Email: Zhaorui.Tan21@student.xjtlu.edu.cn

Oct, 2024

NEURAL INFORMATION PROCESSING SYSTEMS

[*]Corresponding authors

Table of Contents

Generalization settings for visual classification

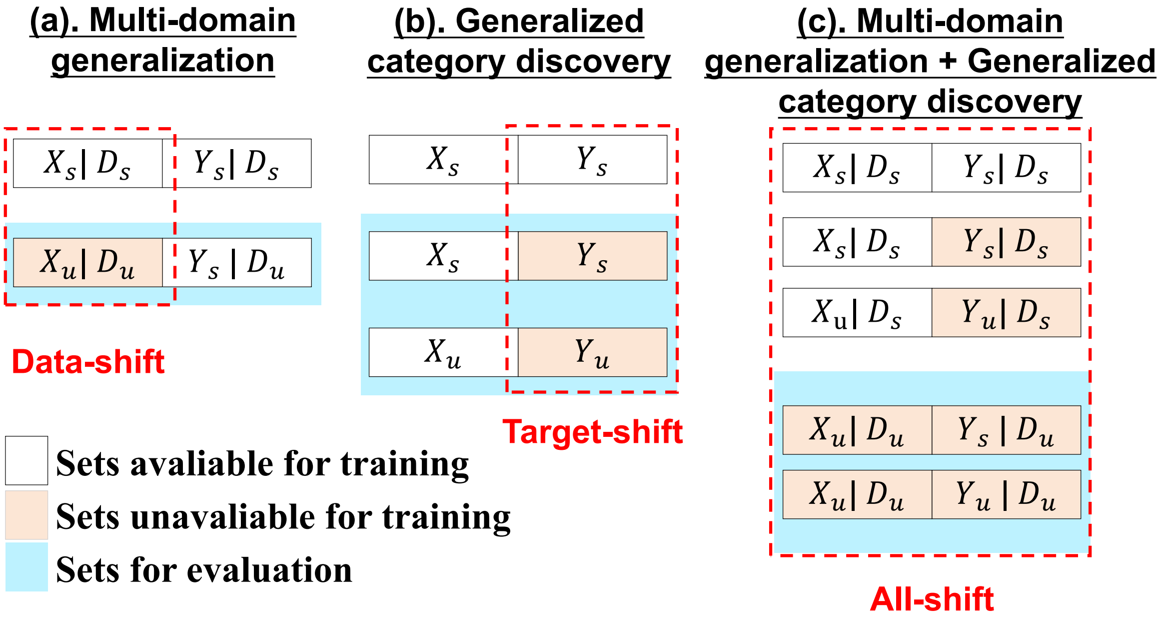


Figure 1: Diagrams of different generalization settings in visual classification tasks.

How can we improve generalization for all these settings?
Can we even improve the interpretability with generalization?

Interpretability and generalization in one: L-Reg

Facing the above questions, we introduce Logic regularization (L-Reg)

$$L_{L-Reg} = -\frac{1}{M}\sum_{i=1}^{M}\left[\sum_{j=1}^{K}\sigma_{j,i}(\hat{Y}^T Z)\log\sigma_{j,i}(\hat{Y}^T Z)\right] + \sum_{j=1}^{K}\left[\frac{1}{M}\sum_{i=1}^{M}\sigma_{j,i}(\hat{Y}^T Z)\log\left(\frac{1}{M}\sum_{i=1}^{M}\sigma_{j,i}(\hat{Y}^T Z)\right)\right], \quad (1)$$

where $\sigma_{j,i}(\hat{Y}^T Z)$ denotes the value at the i, j position of $softmax(\hat{Y}^T Z)$ and the soft-max function is applied at the last dimension.

L-Reg improves generalization with interpretability.

What can L-Reg do? Improving interpretability

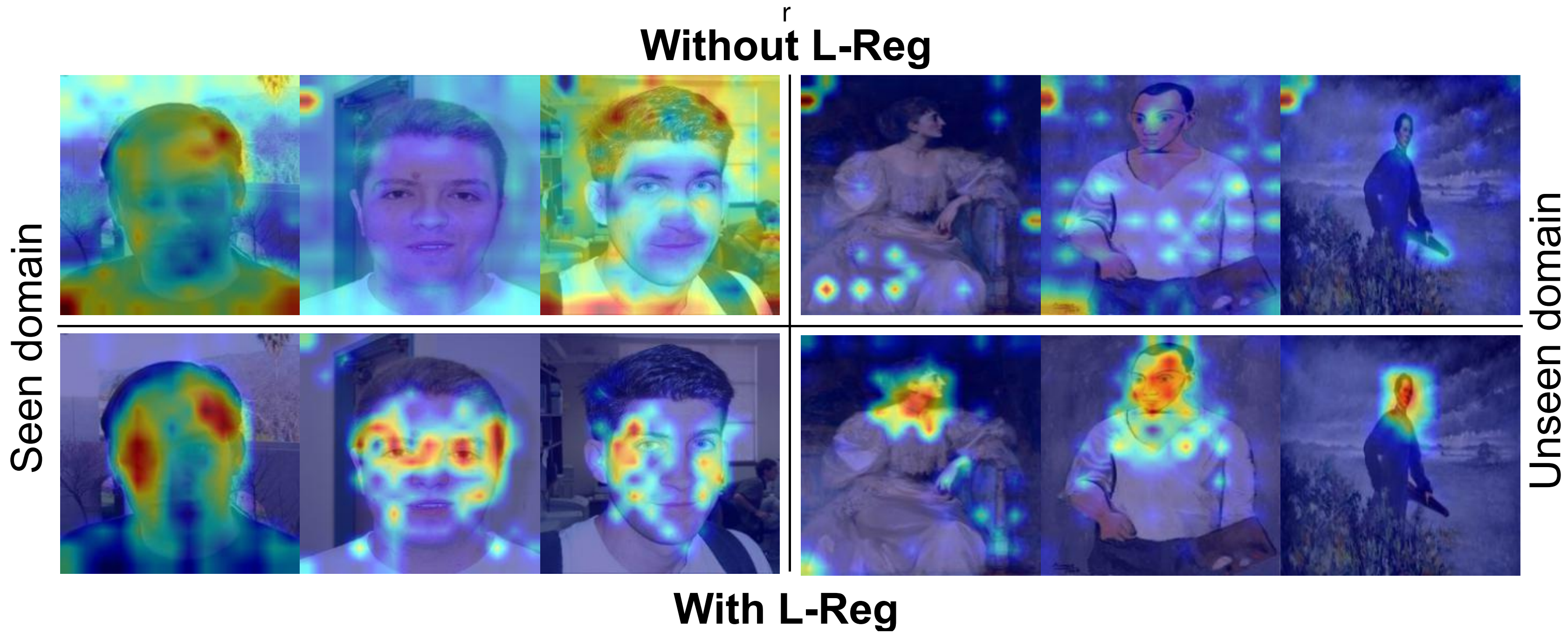
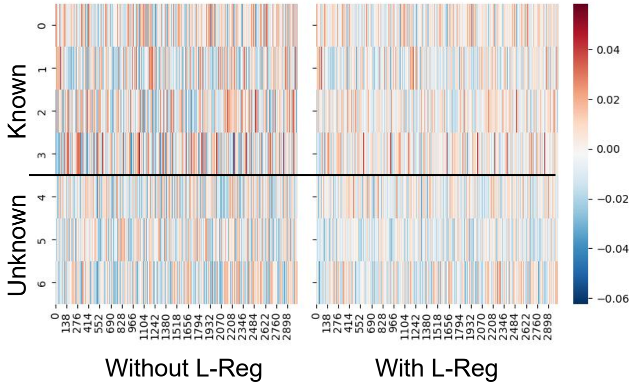
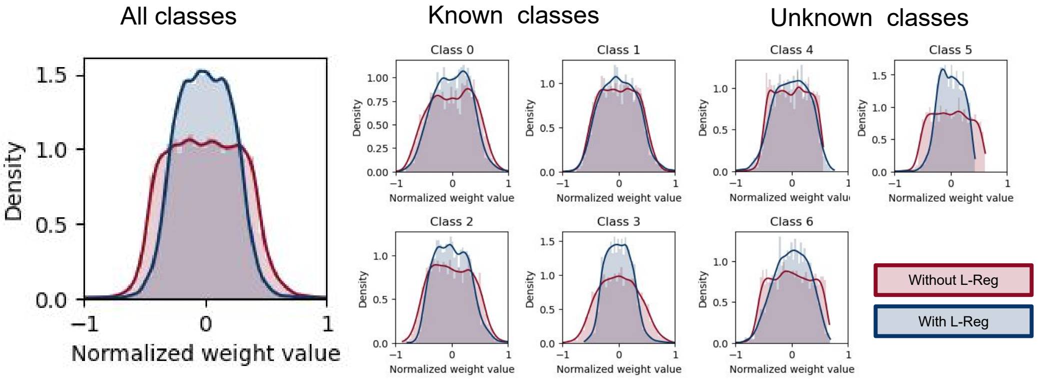


Figure 2: GradCAM [1] visualizations for the unknown class 'person' across seen and unseen domains of the GMDG baseline with L_2 regularization that is trained without and with L-Reg, respectively. Both experiments share the same hyper-parameters, except the latter is using the L-Reg.

What can L-Reg do? Reducing classifier complexity



(a). Heatmap of classifer's weights



(b). Distribution of values of classifer's weights under classes

Figure 3: Visualizations of classifiers' weights form models trained using GMDG on PACS dataset without and with L-Reg under mDD+GCD setting, respectively. Both experiments share the same hyper-parameters using Regnety-16g backbone, except the latter uses additional L-Reg.

What can L-Reg do? Balancing feature complexity

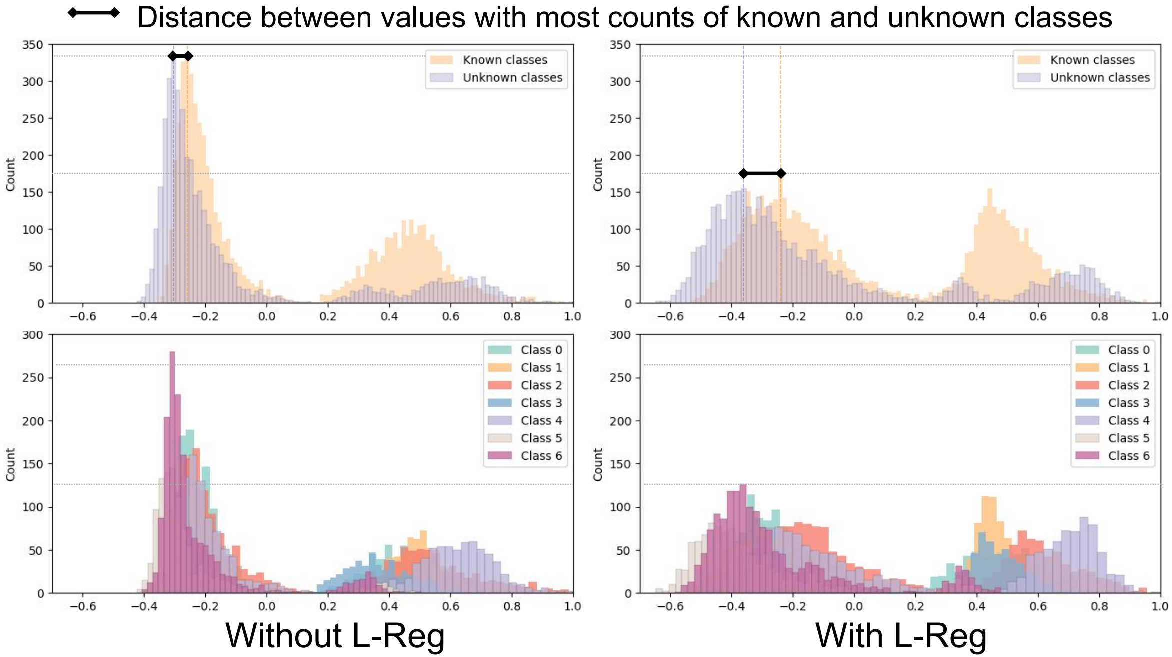


Figure 4: Visualizations of latent features form models trained using GMDG on PACS dataset without and with L-Reg under mDD+GCD setting using RegNetY-16G backbone, respectively.

Logical analysis framework v.s. visual classification task

Definition

Following [2], a logic $\mathcal{L}$ is a five-tuple defined in the form:

$$\mathcal{L} = \langle F_{\mathcal{L}}, M_{\mathcal{L}}, \models_{\mathcal{L}}, mng_{\mathcal{L}}, \vdash_{\mathcal{L}}, \rangle . \tag{2}$$

- $F_{\mathcal{L}}$: a set of all formulas of $\mathcal{L}$. Images and labels (X, Y) for computer vision cases.
- $M_{\mathcal{L}}$: a class called the class of all models (or possible worlds) of $\mathcal{L}$. Different domains D of X.
- $\models_{\mathcal{L}}$: a binary relation, $\models_{\mathcal{L}} \subseteq M_{\mathcal{L}} \times F_{\mathcal{L}}$, called the validity relation of $\mathcal{L}$. In the known set, the ground truth label of the image is given as truth, which is the validity relation.
- $mng_{\mathcal{L}} : F_{\mathcal{L}} \times M_{\mathcal{L}} \longrightarrow$ Sets where Sets is the class of all sets. $mng_{\mathcal{L}}$ is a function with domain $F_{\mathcal{L}} \times M_{\mathcal{L}}$, called the meaning function of $\mathcal{L}$: Classifiers.
- $\vdash_{\mathcal{L}}$ represents the provability relation of $\mathcal{L}$, telling us which formulas are ‘true’ in which possible world and usually is definable from $mng_{\mathcal{L}}$. Estimation criteria.

We can correlate the image classification procedure in computer vision with the framework of logic studies perfectly :)

'Good general' logic

Following Definition 1, on the given X, Y sets, we specify:

$$\mathcal{L}_{(X_s,Y_s)} = \left\langle F_{(X_s,Y_s)}, D, \models_{(X_s,Y_s)}, h, \vdash_{(h(X),Y)} \right\rangle . \tag{3}$$

We aim to achieve a good general logic $\mathcal{L}^*$ from $\mathcal{L}_{(X_s,Y_s)}$ because:

- A *good general* logic has strong generalizability.

By definition, we know that:

- $F_{(g(X_s),Y_s)}$ and h in $\mathcal{L}^*$ should form the *atomic formulas* to achieve the good general logic.

How to form the atomic formulas?

Semantic support

Definition (Semantic support)

We denote $z = g(x)$, where $z \in Z$, as a set of compositions of these semantics: $z := \{z^i\}_{i=1}^{M}$, where M is the number of dimensions or semantics. Notably, not all semantics in z may be useful for deduction or inference. We define the subset γ of z, extracted from the sample $x \sim \mathcal{X}$, as the semantic support of x if γ is sufficient for deducing the relationship between x and a $y \sim \mathcal{Y}$.

Semantic supports gained in latent features combining with the classifier from the atomic formulas: $h(g(x), y, d) \rightarrow True/False, s.t., \vdash_{(h \circ g(X), Y)} = \models_{(g(X_s), Y_s)}$.

Based on the definition of good general logic, we present the constraints of learning semantic supports:

$$\min_{h,g} H(Y|g(\Gamma), D) - H(Y|g(\bar{\Gamma}), D), \tag{4}$$

which derives into Eq.1 as L-Reg.

Results: MDG results

Table 1: MDG results: Comparison between the proposed and previous non-ensemble and ensemble mDG methods. The best results for each group are highlighted in **bold**. Improvement and degradation in our approach from GMDG are highlighted in red.

Test domain	**PACS**	**VLCS**	**OfficeHome**	**TerraIncognita**	**DomainNet**	**Avg.**
MMD [3]	84.7±0.5	77.5±0.9	66.3±0.1	42.2±1.6	23.4±9.5	58.8
Mixstyle [4]	85.2±0.3	77.9±0.5	60.4±0.3	44.0±0.7	34.0±0.1	60.3
GroupDRO [5]	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
IRM [6]	83.5±0.8	78.5±0.5	64.3±2.2	47.6±0.8	33.9±2.8	61.6
ARM [7]	85.1±0.4	77.6±0.3	64.8±0.3	45.5±0.3	35.5±0.2	61.7
VREx [8]	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
CDANN [9]	82.6±0.9	77.5±0.1	65.8±1.3	45.8±1.6	38.3±0.3	62.0
DANN [10]	83.6±0.4	78.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	62.6
RSC [11]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
MTL [12]	84.6±0.5	77.2±0.4	66.4±0.5	45.6±1.2	40.6±0.1	62.9
MLDG [13]	84.9±1.0	77.2±0.4	66.8±0.6	47.7±0.9	41.2±0.1	63.6
Fish [14]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
ERM [15]	84.2±0.1	77.3±0.1	67.6±0.2	47.8±0.6	44.0±0.1	64.2
SagNet [16]	86.3±0.2	77.8±0.5	68.1±0.1	48.6±1.0	40.3±0.1	64.2
SelfReg [17]	85.6±0.4	77.8±0.9	67.9±0.7	47.0±0.3	42.8±0.0	64.2
CORAL [18]	86.2±0.3	78.8±0.6	68.7±0.3	47.6±1.0	41.5±0.1	64.5
mDSDI [19]	86.2±0.2	79.0±0.3	69.2±0.4	48.1±1.4	42.8±0.1	65.1
Use RegNetY-16GF [20] as oracle model.						
MIRO [21] (ECCV23)	97.4±0.2	79.9±0.6	80.4±0.2	58.9±1.3	53.8±0.1	74.1
GMDG [22] (CVPR24)	97.3±0.1	82.4±0.6	80.8±0.6	60.7±1.8	54.6±0.1	75.1
GMDG + L-Reg	**97.4**±0.2 $^{0.1\uparrow}$	**82.4**±0.0 $^{0.1\uparrow}$	**80.9**±0.5 $^{0.1\uparrow}$	**62.9**±0.9 $^{2.2\uparrow}$	**55.3**±0.0 $^{0.8\uparrow}$	**75.8** $^{0.7\uparrow}$

Results: GCD results, mDG+GCD results

Table 2: GCD results: Average results across all datasets of PIM with L-Reg. Improvements and degradation are highlighted in red and blue, respectively.

Average	All	Known	Unknown
K-means [23]	44.7	46.0	43.9
RankStats+ [24] (TPAMI-21)	38.6	54.6	25.6
UNO+ [25] (ICCV-21)	51.2	74.5	36.7
ORCA [26] (ICLR-22)	46.3	51.3	41.2
ORCA - ViTB16	56.7	65.6	49.9
GCD [27] (CVPR-22)	60.4	71.8	52.9
RIM [28] (NeurIPS-10)	62.0	72.5	55.4
TIM [29] (NeurIPS-20)	62.7	72.6	56.4
PIM [30] (ICCV-23)	67.4	**79.3**	59.9
PIM + L-Reg	**68.8** $^{1.4\uparrow}$	79.0 $^{0.3\downarrow}$	**62.7** $^{2.8\uparrow}$

Table 3: **Results of Congestion prediction:** Congestion prediction is proposed for circuit design.

	pearson	spearman	kendall
Gpdl with UNet++	0.6085	0.5202	0.3855
CircuitFormer (SOTA)	0.6374	0.5282	0.3935
CircuitFormer + L-Reg (Ours)	**0.6553**	**0.5289**	**0.3944**

Table 4: MDG+GCD results: Averaged accuracy scores for all, known and unknown classes across all five datasets. Improvements and degradation are highlighted in red and blue respectively.

Method	Domain gap	All	Known	Unknown
ERM	Not	44.69	59.33	23.54
+L-Reg	minimized	45.50	61.43	21.63
Imp.		0.81	2.09	-1.91
PIM	Not	46.95	60.35	26.90
+L-Reg	minimized	47.27	60.83	26.34
Imp.		0.32	0.48	-0.57
MIRO	Not sufficiently	49.67	68.86	25.79
+L-Reg	minimized	52.11	71.26	26.49
Imp.		2.44	2.39	0.71
GMDG		47.94	68.75	20.68
+L-Reg	Minimized	51.94	69.87	27.68
Imp.		4.00	1.12	7.01

Advantages and limitations

L-Reg forms atomic formulas and improves interpretability.

- For known classes:
  - $h(\text{has fingerboard}, \text{is guitar}, d \in D) \to \text{True}$
  - $h(\text{not has fingerboard}, \text{is guitar}, d \in D) \to \text{False}$
- For unknown classes:
  - $h(\text{has a face}, \text{is person}, d \in D) \to \text{True}$
  - $h(\text{not has a face}, \text{is person}, d \in D) \to \text{False}$

Limitations

- It may fail when the domain shift is enormous, e.g., sketch domain where human faces are missing and others.
- One crucial precondition highlighted in the theoretical analysis is that L-Reg operates effectively with a representation Z, where each **dimension represents independent semantics.**

Table 5: Averaged results of applying L-Reg to different layers across domains in PACS.

	All	Known	Unkown
GMDG	58.33	91.46	10.18
L-Reg: Deep layer	**67.82**	**91.86**	31.33
L-Reg: Earlier and the deep layers	58.97	80.73	**35.05**

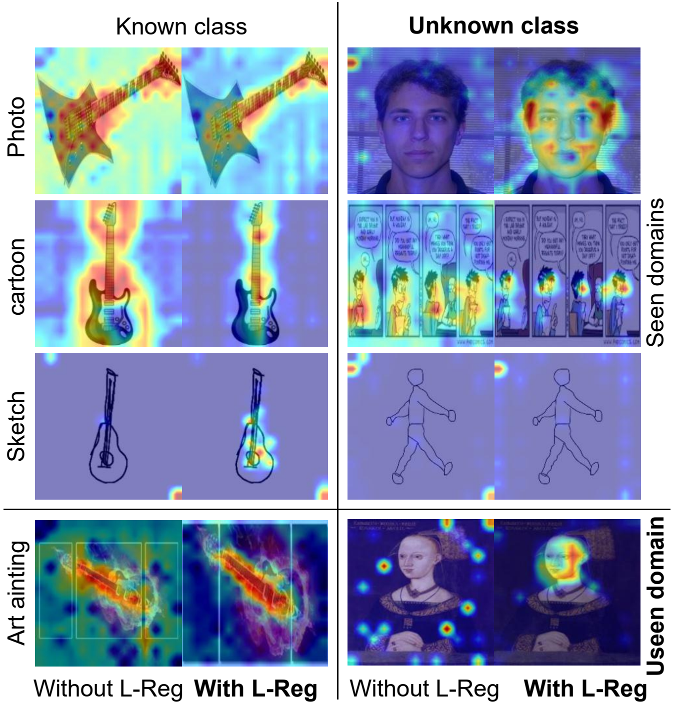


Figure 5: GradCAM visualizations of GMDG trained without and with L-Reg. The seen, unseen domains and known, unknown classes are denoted.

Future work

We provide several possible solutions to the limitations of L-Reg.

- L-Reg should be applied to features from deep layers.
- Constraining the independence of dimensions in Z (e.g., using Ortho-Reg).

Table 6: **Results of GCD:** Averaged results across all datasets of PIM with different regularization applied to the latent features: Sparsity: achieved through Bernoulli Sample; Ortho-Reg: orthogonality regularization. +L-Reg outperforms other regularization terms when they are applied solely; +L-Reg+Ortho-Reg achieves the best performance and alleviates the performance degradation of unknown classes, validating our hypothesis in the paper that the improper Z may result in compromises and constraining the independence of each $z^i \in z, z \in Z$ may be helpful.

	Avg		
	All	Known	Unknown
PIM	67.4	79.3	59.9
+Sparsity	66.6	77.3	60.0
Improvements	-0.7	-2.0	0.1
+Ortho-Reg	68.4	79.2	61.9
Improvements	1.0	-0.1	2.0
+L-Reg	68.8	79.0	62.7
Improvements	1.4	-0.3	2.8
+L-Reg+Ortho-Reg	**69.3**	**79.6**	**63.4**
Improvements	**2.0**	**0.3**	**3.5**

Table 7: **Results of GCD:** Detailed results across all datasets of PIM with different regularization applied to the latent features: Sparsity: achieved through Bernoulli Sample; Ortho-Reg: orthogonality regularization.

	CUB			**Stanford Cars**			**Herbarium19**		
	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
PIM	62.7	75.7	56.2	43.1	66.9	31.6	42.3	56.1	34.8
PIM + Sparsity	60.1	72.7	53.8	40.4	61.7	30.1	42.0	53.7	35.8
Improvements	-2.6	-3.0	-2.4	-2.7	-5.2	-1.5	-0.3	-2.4	1.0
PIM + Ortho-Reg	64.9	76.7	58.9	44.3	65.6	34.1	42.9	57.2	35.1
Improvements	2.2	1.0	2.7	1.2	-1.3	2.5	0.6	1.1	0.3
PIM + L-Reg	65.3	76.0	60.0	44.8	66.0	34.6	**43.7**	55.8	**37.2**
Improvements	2.6	0.3	3.8	1.7	-0.9	3.0	1.4	-0.3	2.4
PIM + L-Reg + Ortho-Reg	**66.8**	**77.3**	**61.6**	**45.8**	**67.3**	**35.5**	43.3	57.5	35.6
Improvements	4.1	1.6	5.4	2.7	0.4	3.9	1.0	1.4	0.8

	CIFAR10			**CIFAR100**			**ImageNet-100**		
	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
PIM	94.7	97.4	93.3	78.3	84.2	66.5	83.1	**95.3**	77.0
PIM + Sparsity	94.2	97.4	92.6	79.7	**84.6**	69.7	83.4	93.7	78.2
Improvements	-0.5	0.0	-0.7	1.4	0.4	3.2	0.3	-1.6	1.2
PIM + Ortho-Reg	95.1	97.4	93.9	80.2	**84.6**	71.4	83.0	93.4	77.7
Improvements	0.4	0.0	0.6	1.9	0.4	4.9	-0.1	-1.9	0.7
PIM + L-Reg	94.8	**97.6**	93.4	80.8	**84.6**	73.2	83.4	94.0	78.0
Improvements	0.1	0.2	0.1	2.5	0.4	6.7	0.3	-1.3	1.0
PIM + L-Reg + Ortho-Reg	**95.1**	**97.6**	**93.9**	**81.2**	84.2	**75.0**	**83.7**	93.6	**78.7**
Improvements	0.4	0.2	0.6	2.9	0.0	8.5	0.6	-1.7	1.7

Wait! You may also be interested in ...

Our previous studies in generalization:

- Multi-domain generalization from statistical perspective:
  Rethinking Multi-domain Generalization with A General Learning Objective (CVPR24). [22]
- An augmentation framework for enhancing generalization in text2image generation that based on group theory:
  Semantic-Aware Data Augmentation for Text-to-Image Synthesis (AAAI24). [31]

Reference I

[1] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: **Proceedings of the IEEE international conference on computer vision**. 2017, pp. 618–626.

[2] Hajnal Andréka, István Németi, and Ildikó Sain. "Universal algebraic logic". In: **Studies in Logic, Springer, due to** (2017).

[3] Haoliang Li et al. "Domain generalization with adversarial feature learning". In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2018, pp. 5400–5409.

[4] Kaiyang Zhou et al. "Domain generalization with mixstyle". In: **arXiv preprint arXiv:2104.02008** (2021).

[5] Shiori Sagawa et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization". In: **arXiv preprint arXiv:1911.08731** (2019).

[6] Martin Arjovsky et al. "Invariant risk minimization". In: **arXiv preprint arXiv:1907.02893** (2019).

[7] Marvin Zhang et al. "Adaptive risk minimization: Learning to adapt to domain shift". In: **Advances in Neural Information Processing Systems** 34 (2021), pp. 23664–23678.

[8] David Krueger et al. "Out-of-distribution generalization via risk extrapolation (rex)". In: **International Conference on Machine Learning**. PMLR. 2021, pp. 5815–5826.

[9] Ya Li et al. "Deep domain generalization via conditional invariant adversarial networks". In: **Proceedings of the European conference on computer vision (ECCV)**. 2018, pp. 624–639.

[10] Yaroslav Ganin et al. "Domain-adversarial training of neural networks". In: **The journal of machine learning research** 17.1 (2016), pp. 2096–2030.

[11] Zeyi Huang et al. "Self-challenging improves cross-domain generalization". In: **Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16**. Springer. 2020, pp. 124–140.

Reference II

[12] Gilles Blanchard et al. "Domain generalization by marginal transfer learning". In: **The Journal of Machine Learning Research** 22.1 (2021), pp. 46–100.

[13] Da Li et al. "Learning to generalize: Meta-learning for domain generalization". In: **Proceedings of the AAAI conference on artificial intelligence**. Vol. 32. 1. 2018.

[14] Yuge Shi et al. "Gradient matching for domain generalization". In: **arXiv preprint arXiv:2104.09937** (2021).

[15] Vladimir N. Vapnik. **Statistical Learning Theory**. Wiley-Interscience, 1998.

[16] Hyeonseob Nam et al. "Reducing domain gap by reducing style bias". In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2021, pp. 8690–8699.

[17] Daehee Kim et al. "Selfreg: Self-supervised contrastive regularization for domain generalization". In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. 2021, pp. 9619–9628.

[18] Baochen Sun and Kate Saenko. "Deep coral: Correlation alignment for deep domain adaptation". In: **Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14**. Springer. 2016, pp. 443–450.

[19] Manh-Ha Bui et al. "Exploiting domain-specific features to enhance domain generalization". In: **Advances in Neural Information Processing Systems** 34 (2021), pp. 21189–21201.

[20] Mannat Singh et al. "Revisiting weakly supervised pre-training of visual perception models". In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2022, pp. 804–814.

[21] Cha Junbum et al. "Domain Generalization by Mutual-Information Regularization with Pre-trained Models". In: **European Conference on Computer Vision (ECCV)** (2022).

Reference III

[22] Zhaorui Tan, Xi Yang, and Kaizhu Huang. "Rethinking Multi-domain Generalization with A General Learning Objective". In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. June 2024, pp. 23512–23522.

[23] J MacQueen. "Classification and analysis of multivariate observations". In: **Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability**. 1967, pp. 281–297.

[24] Kai Han et al. "Autonovel: Automatically discovering and learning novel visual categories". In: **IEEE Transactions on Pattern Analysis and Machine Intelligence** (2021).

[25] Enrico Fini et al. "A unified objective for novel class discovery". In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. 2021, pp. 9284–9292.

[26] Kaidi Cao, Maria Brbic, and Jure Leskovec. "Open-World Semi-Supervised Learning". In: **International Conference on Learning Representations**. 2022. URL: https://openreview.net/forum?id=O-r8LOR-CCA.

[27] Sagar Vaze et al. "Generalized category discovery". In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2022, pp. 7492–7501.

[28] Andreas Krause, Pietro Perona, and Ryan Gomes. "Discriminative clustering by regularized information maximization". In: **Advances in neural information processing systems** 23 (2010).

[29] Malik Boudiaf et al. "Information maximization for few-shot learning". In: **Advances in Neural Information Processing Systems** 33 (2020), pp. 2445–2457.

[30] Florent Chiaroni et al. "Parametric information maximization for generalized category discovery". In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. 2023, pp. 1729–1739.

Reference IV

[31] Zhaorui Tan, Xi Yang, and Kaizhu Huang. "Semantic-Aware Data Augmentation for Text-to-Image Synthesis". In: **Proceedings of the AAAI Conference on Artificial Intelligence**. Vol. 38. 6. 2024, pp. 5098–5107.

Thank You !