Interpret Your Decision: Logical Reasoning Regularization for Generalization in Visual Classification

Zhaorui Tan[1,2], **Xi Yang**[1]*, **Qiufeng Wang**[1], **Anh Nguyen**[2], **Kaizhu Huang**[3]*
[1] Xi'an-Jiaotong Liverpool University
[2] University of Liverpool
[3] Duke Kunshan University

Abstract

Vision models excel in image classification but struggle to generalize to unseen data, such as classifying images from unseen domains or discovering novel categories. In this paper, we explore the relationship between logical reasoning and deep learning generalization in visual classification. A logical regularization termed L-Reg is derived which bridges a logical analysis framework to image classification. Our work reveals that L-Reg reduces the complexity of the model in terms of the feature distribution and classifier weights. Specifically, we unveil the interpretability brought by L-Reg, as it enables the model to extract the salient features, such as faces to persons, for classification. Theoretical analysis and experiments demonstrate that L-Reg enhances generalization across various scenarios, including multi-domain generalization and generalized category discovery. In complex real-world scenarios where images span unknown classes and unseen domains, L-Reg consistently improves generalization, highlighting its practical efficacy.

1 Introduction

One critical challenge in visual classification models is their ability to generalize effectively to unseen samples or unknown classes. For instance, a model trained on real images of various animals should ideally classify animal sketches accurately (referred to as multi-domain generalization classification [20, 35, 34, 23, 25, 37, 50]) or discover novel categories not present in the training set (referred to as generalized category discovery [54, 16]). These problems are prevalent in real-world scenarios, where training data-

Figure 1: GradCAM [45] 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 uses the L-Reg.

target pairs are usually insufficient, and labeling is time-consuming so that not every data is paired with a label. Meanwhile, test data is likely to contain shifts in both data and targets, making it essential to propose methods that generalize to border scenarios.

Regularization terms, such as L_2 regularization leading to weight decay, are commonly employed during training to improve a model's generalization capabilities. However, the L_2 regularization
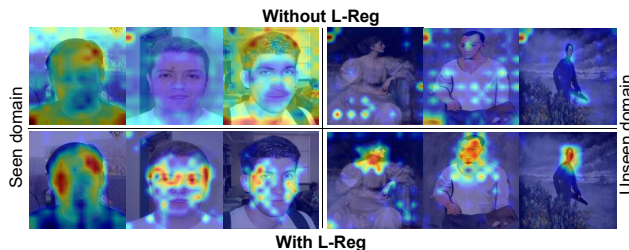
---

*Corresponding authors. <Xi.Yang01@xjtlu.edu.cn; kaizhu.huang@dukekunshan.edu.cn>

is *parametric-based* rather than *sample-based*, which may lead to ambiguous interpretability [58]. As illustrated in Fig. 1, the model trained solely with L_2 regularization exhibits low interpretability. Other regularization terms [57, 58, 59] attempt to improve the interpretability of deep learning models for sequential signals rather than vision, whereas [39] proposes a regularization term to enhance interpretability for robustness in visual classification models rather than generalization. Drawing inspiration from logical reasoning has shown promise for better generalization and interpretability in various tasks. Current work unveils the effectiveness of logical reasoning in generalization tasks, such as boosting performance in length generalization [1, 3, 2, 60] and abstract symbol relational reasoning [10, 36] (e.g., mathematical solving and psychological tests). Several efforts, such as [6], explore the explicit entropy-based logical explanations of neural networks for image classification, confirming the presence and interpretability of logical reasoning within visual tasks. Yet, there are limited studies tackling the generalization of visual classification tasks through the lens of logical reasoning.

This paper studies two pivotal questions corresponding to the above: *1) How does logical reasoning relate to visual tasks such as image classification? 2) How can we derive a logical reasoning-based regularization term to benefit generalization?* To achieve these, we correlate the image classification procedure in computer vision with the framework of logic studies [4], positing that training an image classifier involves learning a *good general* logical relationship between images and labels via an encoder. This good general logic is attained when the semantics generated by the encoder and classifier can be combined to form atomic formulas. Our exploration leads to the introduction of a sample-based Logical regularization term named L-Reg. We reveal that L-Reg efficiently reduces the *complexity* of the model from two aspects: 1) L-Reg leads to a balanced feature distribution in the semantic space; 2) L-Reg reduces the number of weights with extreme values in the classifier.

Intuitively, the complexity reduction achieved by L-Reg stems from its ability to filter out redundant features or semantics, focusing instead on the minimal yet sufficient semantics for classification - defined as semantic support in Definition 3.2, where the interpretability also emerges. This filtering feature benefits the generalization when there is a domain shift in data where the domain-dependent features are ignored for classification. Moreover, it further promotes generalization when unlabeled data from the unknown classes is present. If such data lacks the semantic support associated with known classes, it is then classified as belonging to an unknown class, and its corresponding semantic supports are extracted. These capabilities equip L-Reg with explicit interpretability. As Fig. 1 shows, with L-Reg, the model can identify the unknown class 'person', and pinpoint faces which are the crucial features for classifying this category. In contrast, the model trained solely with L_2 (without L-Reg) focuses on the ambiguous features for classification.

Rigorous theoretical analysis and experimental results validate that L-Reg yields better generalization across diverse scenarios. Specifically, L-Reg facilitates better performance under the aforementioned multi-domain generalization and generalized category discovery tasks, whose settings are presented in Fig. 2 (a)(b). Furthermore, to evaluate L-Reg's robustness, we introduce a more complex real-world scenario, as shown in Fig. 2 (c), where unlabeled images may not only belong to unknown classes, but also originate from unseen domains. Even in this challenging context, L-Reg is still able to consistently demonstrate notable improvements in generalization, underscoring its practical utility and effectiveness. Our code is available at `https://github.com/zhaorui-tan/L-Reg_NeurIPS24`.

2 Preliminaries and generalization settings for visual classification

Consider paired $(X, Y) \sim (\mathcal{X}, \mathcal{Y})$, $(X_s, Y_s) \sim (\mathcal{X}_s, \mathcal{Y}_s)$, and $(X_u, Y_u) \sim (\mathcal{X}_u, \mathcal{Y}_u)$ denote all sets of inputs and labels, seen paired subsets of (X, Y), and unseen paired subsets of (X, Y), respectively. Note that X_u, Y_u may be accessible for the model separately, but their pairing relationships are not accessible. Let D denote the possible domains, with $D_s, D_u \subset D$ representing the seen and unseen domains. In classification tasks, an encoding function $g(x) \rightarrow Z \in \mathbb{R}^M$ is commonly introduced to map X into the latent feature set Z, where each latent feature has M dimensions. A predictor $h(Z) \rightarrow \hat{Y} \in \mathbb{R}^K$ maps Z to predictions $\hat{Y}$, where K denotes
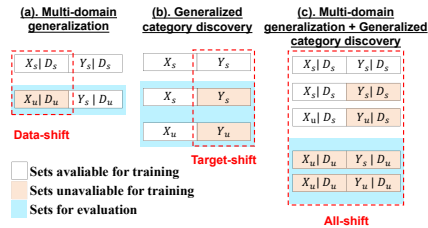

Figure 2: Diagrams of different generalization settings in visual classification tasks.
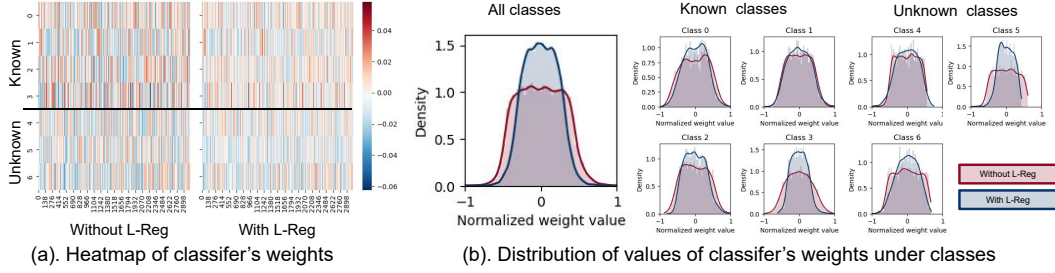
Figure 3: Visualizations of classifiers' weights form models trained using GMDG on PACS dataset without and with L-Reg under mDG+GCD setting, respectively. Both experiments share the same hyper-parameters using Regnety-16g backbone, except the latter uses additional L-Reg.

the number of classes and the dimensions of predictions. $P(\cdot)$ and $H(\cdot)$ symbolize probability and entropy, respectively. This paper discusses two typical cases for generalization in image classification tasks: (1) *Data-shift generalization:* X_s and X_u have distribution shifts, such as multi-domain generalization (mDG); and (2) *Target-shift generalization:* Y_s and Y_u have distribution shifts, which stands for tasks like generalized category discovery (GCD). We additionally explore a challenging scenario called *All-shift generalization:* both X_s and X_u, Y_s and Y_u have distribution shifts, which is a combination of mDG and GCD tasks (mDG + GCD). The following lists the detailed settings for generalization. Please refer to Fig. 2 for brief diagrams.

Data-shift generalization: Problem setting for mDG. Illustrated in Fig. 2 (a), mDG [9] intends to generalize well to unseen domains having the objective of $\min H(X_s, Y_s \mid D_s)$ and expecting the model to be generalized to X_u when predicting Y_u from the unseen domain D_u. In such cases, Y_u is fully accessible to the model since Y_s and Y_u share the same domain: $\mathcal{Y}_s = \mathcal{Y}_u$ but there are shifts in X where $\mathcal{X}_s \neq \mathcal{X}_u$.

Target-shift generalization: Problem setting for GCD. GCD [54] (Fig. 2 (b)) aims to discover possible unseen labels among unlabeled datasets X_u. The challenge is that the samples in X_u may belong to known classes or unknown classes: $\mathcal{Y}_s \neq \mathcal{Y}_u$ and probably $\mathcal{Y}_s \cap \mathcal{Y}_u \neq \emptyset$. The model should be able to distinguish the samples from the known classes and cluster the samples for unknown classes simultaneously. Note that X_u is used for model training, but the relationship between X_u and Y_u is unseen for the model. In summary, shifts exist between Y_s and Y_u but not between X_s and X_u.

All-shift generalization: Problem setting for mDG + GCD. To explore the generalization problem further, we introduce a setting that is the combination of mDG and GCD as shown in Fig. 2 (c). Specifically, the model is trained on the labeled pairs (X_s, Y_s) and unlabeled set X_u from the seen domains D_s; X_u may belong to known and unknown classes. Furthermore, the model is tested on X_u from the unseen domain D_u, where X_u may also come from the known and unknown classes. In this setting, the model is expected to 1) classify samples to the seen classes and discover the unseen classes among unlabeled samples from seen domains and 2) generalize this ability to the samples from the unseen domain. In this scenario, X_s and X_u have shifts, and so do Y_s and Y_u.

For all aforementioned generalization settings, the objective can be summarized as minimizing the *generalization loss*:

Definition 2.1 (Generalization loss). Let the target model $f^* : f^*(X, Y) : X \rightarrow Y$, can generalize across both seen and unseen sets X, Y. Denote its trainable f, which is only trained on the seen sets. The generalization loss for the unseen sets is defined as:

$$GL(f, f^*, (X_u, Y_u)) = \mathbb{E}_{(x,y)\in(X_u,Y_u)}||f(x,y) - f^*(x,y)||_2. \tag{1}$$

3 Logical regularization for generalization in image classification

Under the problem settings defined in Section 2, we introduce Logic regularization (L-Reg) targeting the objective:

$$\min_{h,g} \mathbb{E}_{z_i \in z, z \in Z}[H(\hat{Y}|z_i, D)] - \mathbb{E}_{z \in Z}[H(\hat{Y}|Z, D)], \tag{2}$$

3

where $\hat{Y} \in \mathbb{R}^K = h \circ g(X)$ is the prediction set. The corresponding Logic regularization loss (L-Reg) is defined as:

$$L_{L-Reg} = \frac{1}{M} \sum_{i=1}^{M} \left[\sum_{j=1}^{K} [\sigma(\hat{Y}_j^T Z_i) \log \sigma(\hat{Y}_j^T Z_i)] - [\frac{1}{K} \sum_{j=1}^{K} \sigma(\hat{Y}_j^T Z_i) \log(\frac{1}{K} \sum_{j=1}^{K} \sigma(\hat{Y}_j^T Z_i)] \right], \quad (3)$$

where $\sigma(\hat{Y}_j^T Z_i)$ 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. By incorporating other existing methods' losses denoted by L_{main}, the overall loss is formulated as:

$$L_{all} = L_{main} + \alpha L_{L-Reg}, \quad (4)$$

with a weight α applied to balance two losses. As depicted in Fig. 1, L-Reg plays a pivotal role in extracting crucial features for image classification, thus enhancing generalization capabilities. This beneficial outcome can be attributed to two primary factors:

Reducing classifier complexity: L-Reg streamlines the complexity of the classifier itself, as depicted in Fig. 3 (a). Notably, the heat map of the model with L-Reg displays fewer extremely valued weights, evidenced by the diminished presence of intense blue and red colors. This reduction implies that the classifier focuses on leveraging semantically rich and relevant features for decision-making (classification), sidelining the less relevant ones. Additionally, Fig. 3 (b) reveals a reduction in the number of semantic features used to classify each class.

Balancing feature complexity: L-Reg results in a more balanced distribution of features compared to the baseline, as illustrated in Fig. 4. This balanced distribution suggests the elimination of certain extracted semantics characterized by dominant frequencies across all samples. Semantics that occur frequently across samples often lack decisiveness for classification. Hence, reducing their prominence contributes to more expressive feature space and less complex feature distributions. Coupled with the reduced classifier complexity, a simplified classifier achieved through L-Reg facilitates improved generalization across various settings. Specifically, the top row also indicates the distance between the feature distributions of the known and unknown classes, which is enlarged; thus, they are more dividable, leading to classification improvements.
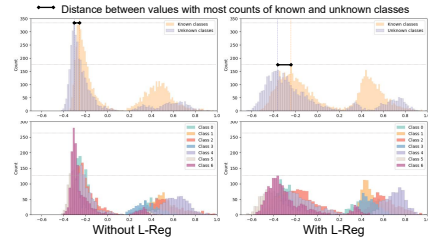
Figure 4: Visualizations of latent features form models trained using GMDG on PACS dataset without and with L-Reg under mDD+GCD setting using RegNetY-16G backbone, respectively.

We present a logical-based theoretical analysis in Section 3.1 and provide the derivation details of L-Reg in Section 3.2. In addition, we discuss the efficacy of L-Reg under various generalization settings in Section 4. Furthermore, L-Reg serves as a plug-and-play loss function that is compatible with most existing frameworks. We conduct experiments applying L-Reg to various established approaches across different generalization settings, as outlined in Section 5.

3.1 Logical framework for visual classification

This part provides the connections between logical reasoning and visual classification tasks. We would like to remind readers of the framework for studying logics and link it with our practical scenarios.

Definition 3.1. Following [4], a logic $\mathcal{L}$ is defined as a five-tuple in the form:

$$\mathcal{L} = \langle F_{\mathcal{L}}, M_{\mathcal{L}}, \models_{\mathcal{L}}, mng_{\mathcal{L}}, \vdash_{\mathcal{L}}, \rangle, \quad (5)$$

where 1) $F_{\mathcal{L}}$ denotes the set of formulas formed by images and labels (X, Y); 2) $M_{\mathcal{L}}$ represents different domains D of X; 3) $\models_{\mathcal{L}}$ is a binary relation relating the truth of whether the formulas are true or false, which has $\models_{\mathcal{L}} \subseteq M_{\mathcal{L}} \times F_{\mathcal{L}}$; 4) $mng_{\mathcal{L}} : F_{\mathcal{L}} \times M_{\mathcal{L}} \longrightarrow$ Sets defines the meaning of X as determined by classifiers, where Sets indicate the class of all sets. (5) $\vdash_{\mathcal{L}}$ symbolizes the provability relation of $\mathcal{L}$, evaluating formulas formed by $mng_{\mathcal{L}}$ is true or false in one possible world, such as the estimation criteria. More details of $\mathcal{L}$ can be seen in Appendix B.

For clarity, we specify $\mathcal{L}_{(X_s,Y_s)} = \left\langle F_{(X_s,Y_s)}, D, \models_{(X_s,Y_s)}, h, \vdash_{(h(X),Y)} \right\rangle$ as the logic formed on the given X, Y sets. With the goal for logic to generalize across a broader scenario and provide extrapolation across all possible formulas in $\mathcal{L}$, a good general logic $\mathcal{L}^*$ should be derived from $\mathcal{L}$ through the feature extractor g:

$$\mathcal{L}^* = \left\langle F_{(g(X_s),Y_s)}, D, \models_{(g(X_s),Y_s)}, h, \vdash_{(h \circ g(X),Y)} \right\rangle, s.t., \vdash_{(h \circ g(X),Y)} = \models_{(g(X_s),Y_s)} . \qquad (6)$$

Importantly, as a good general logic, $F_{(g(X_s),Y_s)}$ and h in $\mathcal{L}^*$ should form the *atomic formulas*, i.e., the tuple of terms with a predicate: $h \circ g(x)$ belongs/not belongs to class y in domain $d \rightarrow Ture/False$, where $x, y, d \in X, Y, D$, which makes that $\vdash_{(h \circ g(X_u),Y_u)} = \models_{(g(X_s),Y_s)}$ still holds. We simply denote one atomic formula in the form of $h(g(x), y, d)$ mapping to binary values. Additionally, $\vdash_{(h \circ g(X),Y)} = \models_{(g(X_s),Y_s)}$ in Eq. (6) can be safely omitted in the rest of the paper. Please see more details about the conditions of the good general logic in Appendix B.

An additional tool is necessary to convert the logic problem into a continuous form, enabling the application of machine learning algorithms. The conditional entropy-based method enables a logically sound derivation of knowledge from the provided dataset with constraints [43]. Specifically, the probabilistic inference process adheres to a probabilistic version of Modus Ponens: $A \rightarrow B, A \vdash B$ (if A then B; not A therefore not B). It is important to note that the logical propositions in probabilistic Modus Ponens are uncertain, with the conditional probability replacing the material implication $A \rightarrow B$. This framework allows us to interpret logical deduction through the lens of entropy. Therefore, for Eq. (6) which implies

$$\exists h \circ g, \ \forall (x,y) \in (X,Y), \ \forall d \in D, \ h \circ g(x) \rightarrow y, \qquad (7)$$

finding $h \circ g$ through optimization is equivalent to

$$\max_{h,g} \mathbb{E}_{(x,y) \in (X,Y), d \in D} P(y|g(x), d) - \mathcal{R} \iff \min_{h,g} \mathbb{E}_{(x,y) \in (X,Y), d \in D} H(y|g(x), d) + \mathcal{R}, \qquad (8)$$

where $\mathcal{R}$ denotes any other possible regularization.

As the logical framework for image classification takes shape, it becomes evident that the unresolved question of identifying an appropriate function g to generate suitable atomic formulas emerges as a critical factor in ensuring the effectiveness of the overarching logic $\mathcal{L}^*$. This paper proposes L-Reg as the regularization to ensure $F_{(g(X_s),Y_s)}$ are formed by atomic formulas in Section 3.2.

3.2 Constructing atomic formulas using L-Reg

In this part, we show the derivation details of L-Reg the aims to ensure the formation of suitable atomic formulas, as depicted in Eq. (6). As highlighted in [1], current algorithms may induce implicit biases towards unseen data, resulting in varied solutions for such data. However, expecting an algorithm to generalize effectively to unseen data domains without appropriate incentivization, such as specifically designed regularization, is unreasonable. Therefore, we aim to enhance the generalization capability of models by employing a logic-based regularization approach. To this end, we introduce the concept of *semantic support* for image classification.

Definition 3.2 (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 $y \sim \mathcal{Y}$.

For instance, if the subset $\{z^1, z^2\} \subseteq z$ is sufficient for accurate inference, the values of other semantics $\{z^i\}_{i=3}^M$ will not impact the inference process. When $\{z^1, z^2\}$ constitutes the minimal combination of semantics required for inference, it is termed the semantic support. We denote Γ as the set of semantic supports of X for deducing each individual class.

Derivation of L-Reg. Regarding Eq. (6), if the semantic supports and their relationship with Y form atomic formulas, Eq. (6) holds as a good general logic, and the generalization would be improved. Thus, we aim to learn the latent features Z, which contain sufficient semantic supports for the deduction of Y:

$$\exists \gamma \in \Gamma, \gamma \subseteq z, \ \forall (z,y) \in (Z,Y), \forall d \in D, \ h(\gamma|d) \rightarrow y. \qquad (9)$$

Specifically, $g(\cdot)$ should meet the following:

$$\forall (\Gamma_i, y_i), (\Gamma_j, y_j) \in (Z,Y), \forall d \in D, \ y_i \neq y_j \iff \Gamma_i \neq \Gamma_j, \qquad (10)$$

5

i.e., the semantic support set for each class should be distinct. The multiple-class classification task has that $\forall \Gamma, |\Gamma| \leq M$. Under the constraints demonstrated in Eq. (9) and Eq. (10), we need to achieve the following through optimization:

$$\min_{h,g} H(Y|g(\Gamma), D), \max_{h,g} H(Y|g(\bar{\Gamma}), D) \iff \min_{h,g} H(Y|g(\Gamma), D) - H(Y|g(\bar{\Gamma}), D), \quad (11)$$

where $\bar{\Gamma}$ denotes the negation of Γ, i.e., the set of semantics which does not include semantic support.

Intuitively, Eq. (11) regularizes that the model should be able to judge whether a sample belongs to a class by using a minimal set of semantic supports; simultaneously, the semantic support sets are also implicitly disentangled for each class, not only for maintaining rich and useful semantics but also for enhancing the independence of deduction of each class. The actual collection of Γ appears to be intractable during optimization. Hence, we resort to deriving its bounds. Regarding Eq. (11), its former term can be elaborated as follows:

$$H(Y|g(\Gamma), D) \leq H(Y|h(z_i), D) \leq \mathbb{E}_{z_i \sim z}[H(Y|g(z_i), D)], \quad (12)$$

where z_i is minimal semantics form z, and $\mathbb{E}_{i=1}^{M} H(Y|g(z_i), D)$ is the upper-bound for $\min_{h,g} H(Y|g(\Gamma), D)$. Therefore, minimizing $\mathbb{E}_{i=1}^{M} H(Y|g(z_i), D)$ is equivalent to minimizing $H(Y|g(\Gamma), D)$. Meanwhile, for the latter in Eq. (11), we have:

$$H(Y|g(\bar{\Gamma}), D) \geq H(Y|g(z), D), \quad (13)$$

where $H(Y|h(z)), D)$ is the lower-bound for $\max_{h,g} H(Y|g(\bar{\Gamma}), D)$. Combining the aforementioned bounds, we have the L-Reg objective as Eq. (2).

Interpretability of semantic supports roots in forming atomic formulas. The atomic formula $\mathcal{A}^y$ is of the form $h(g(x), y, d)$. Our aim is to find the good (most) general $\mathcal{A}^{y*} \in \mathcal{A}^y$ for y class from which the interpretability of L-Reg is derived. Consider $\mathcal{A}_1^y, \mathcal{A}_2^y \in \mathcal{A}^y$, if $\mathcal{A}_1^y$ is more general than $\mathcal{A}_2^y$, there will be a substitution ψ such that $\mathcal{A}_1^y \psi = \mathcal{A}_2^y$ [52]. $\mathcal{A}^{y*}$ should meet $\mathcal{A}^{y*}\psi = \mathcal{A}_i^y \in \mathcal{A}^y$, which infers that $\gamma^y \psi = z^y$ (cf. Eq. (9)) for predication of y where γ^y is the semantic support. Note here that the form of $\mathcal{A}^y$ is constructed for $y \in Y$, i.e., predicate whether the sample belongs to the y class. Considering multiple classes $y_i, y_j \in Y, i \neq j$, it has $\mathcal{A}^{y_i*} \neq \mathcal{A}^{y_j*}$ thus $\gamma^{y_i} \neq \gamma^{y_j}$ (cf. Eq. (10)), which constrains that different minimal semantic supports should be used for predicting different classes. The interpretability of L-Reg is based on $\mathcal{A}^{y*}$, compelling the model to use distinct minimal semantic supports for each class. These minimal semantic supports can be interpreted as the most critical features for efficient prediction. For example, as shown in Fig. 1, the model with L-Reg has learned the facial features of the person class (see more examples in Appendix Figs. 7 to 12), forming the (informal) atomic formula $h(\text{has a human face}, \text{is person}, d \in D) \rightarrow$ True. Similarly, it also leads to $h(\text{not has a human face}, \text{is person}, d \in D) \rightarrow$ False.

4 L-Reg under different generalization settings

L-Reg under data-shift generalization. The task mDG endeavors to facilitate a model's ability to generalize to unseen domains by fostering invariance across seen domains [50]. In the context of mDG, the term $|D| \geq 2$ in Eq. (8) typically denotes multiple domains. Traditionally, existing methods focus on minimizing domain gaps, leading to remarkable results [25, 50]. However, it is noteworthy that even when the domain gap is effectively minimized, and $|D| = 1$ for the latent features can be considered, L-Reg still demonstrates its efficacy in promoting the generalization of X_u from D_u.

Proposition 4.1 (Effectiveness of L-Reg in enhancing data-shift generalization.)**.** *Assume the gap across all domains is well minimized. Let f^* denote the target model that generalizes to the data X_u from the unseen domain with the lowest complexity. For a model $f_{(X_s, Y_s)}^R, f_{(X_s, Y_s)}$ trained under the data-shift generalization setting (i.e., (X_s, Y_s) is accessible and $\mathcal{Y}_s = \mathcal{Y}_u$). We have:*

$$GL(f_{(X_s, Y_s)}^R, f^*, X_u) \leq GL(f_{(X_s, Y_s)}, f^*, X_u). \quad (14)$$

Please see proof details in Proposition C.1. To illustrate Proposition 4.1, consider the following intuitive example: In the seen domains, all cats are either black or white, while all dogs are brown. Now, imagine encountering a sample labeled 'a brown cat' from an unseen domain. Without the application of L-Reg, the model might erroneously classify it as a dog. However, with L-Reg in

6

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 [33]	84.7±0.5	77.5±0.9	66.3±0.1	42.2±1.6	23.4±9.5	58.8
Mixstyle [62]	85.2±0.3	77.9±0.5	60.4±0.3	44.0±0.7	34.0±0.1	60.3
GroupDRO [44]	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
IRM [5]	83.5±0.8	78.5±0.5	64.3±2.2	47.6±0.8	33.9±2.8	61.6
ARM [61]	85.1±0.4	77.6±0.3	64.8±0.3	45.5±0.3	35.5±0.2	61.7
VREx [30]	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
CDANN [35]	82.6±0.9	77.5±0.1	65.8±1.3	45.8±1.6	38.3±0.3	62.0
DANN [20]	83.6±0.4	78.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	62.6
RSC [24]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
MTL [8]	84.6±0.5	77.2±0.4	66.4±0.5	45.6±1.2	40.6±0.1	62.9
MLDG [31]	84.9±1.0	77.2±0.4	66.8±0.6	47.7±0.9	41.2±0.1	63.6
Fish [46]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
ERM [53]	84.2±0.1	77.3±0.1	67.6±0.2	47.8±0.6	44.0±0.1	64.2
SagNet [40]	86.3±0.2	77.8±0.5	68.1±0.1	48.6±1.0	40.3±0.1	64.2
SelfReg [26]	85.6±0.4	77.8±0.9	67.9±0.7	47.0±0.3	42.8±0.0	64.2
CORAL [48]	86.2±0.3	78.8±0.6	68.7±0.3	47.6±1.0	41.5±0.1	64.5
mDSDI [12]	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 [47] as oracle model.				
MIRO [25] (ECCV23)	97.4±0.2	79.9±0.6	80.4±0.2	58.9±1.3	53.8±0.1	74.1
GMDG [50] (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$^{0.7\uparrow}$**

place, the model is compelled to rely on minimal semantics for classification. This means filtering out irrelevant features such as color terms, thus enabling more accurate deductions.

L-Reg under target-shift generalization. We demonstrate how L-Reg enhances generalized discovery in scenarios where only a subset of classes (Y_s) is available for training, and there may exist an overlap between the unseen classes (Y_u) and the seen classes (Y_s), denoted as $\mathcal{Y}_u \cap \mathcal{Y}_s \neq \emptyset$. We define Y_u/Y_s as the novel classes not included in Y_s, and $Y_u \sim Y_s$ as the seen classes for X_u classification, where $|D| = 1$. Building upon Proposition 4.1, L-Reg further enhances GCG by improving the generalization performance on Y_u.

Proposition 4.2 (L-Reg improves target-shift generalization). *When $|D| = 1$, L-Reg promotes generalization performance on Y_u under the target-shift scenario.*

Proof. When $|D| = 1$, since all Y belongs to a close set, minimizing $-H(Y_s|g(\bar{\Gamma}), D)$ is equivalent to the following:

$$\min_{h,g} -H(Y_s|g(\bar{\Gamma})) \iff \min_{h,g} H(\bar{Y}_s|g(\bar{\Gamma})), \tag{15}$$

where $\bar{Y}_s$ is the negation of Y_s, i.e., Y_u/Y_s. In this situation, if one sample does not contain sufficient semantic support to be classified under Y_s, it otherwise will be assigned under Y/Y_s, promoting performance for both Y_u/Y_s and $Y_u \sim Y_s$. Therefore, the generalization performance on the unseen classes will be improved by L-Reg. $\square$

L-Reg under all-shift generalization. When the domain gap is sufficiently minimized and $|D| = 1$ can be considered, the combination of Proposition 4.1 and Proposition 4.2 demonstrates that L-Reg enhances generalization performance on both novel classes (Y_u/Y_s) and seen classes ($Y_u \sim Y_s$) for X_u from other domains. Our experiments validate that L-Reg, when applied in scenarios with well-minimized domain gaps, consistently improves generalization across all shifts.

5 Experiments

To validate L-Reg, three groups of experiments under the three kinds of settings are conducted. Notably, all baselines we used already incorporate the L_2 regulation in the form of weight decay. We also compare other commonly used regularization terms, such as independence or sparsity regularization on Z. More results in Appendix F indicate that our L-Reg also surpasses them.

5.1 Experiments on mDG

Experimental settings. We operate on the DomainBed suite [21] and leverage standard leave-one-out cross-validation as the evaluation protocol. We test L-Reg with GMDG [50] on 5 real-world benchmark datasets: PACS [32], VLCS [18], OfficeHome [55], TerraIncognita [7], and

DomainNet [42]. Following MIRO [25] and GMDG [50], the RegNetY-16GF backbone with SWAG pre-training [47]) is used. Specifically, we train the backbone using GMDG with L-Reg. Accuracy is adopted as the evaluation metric, and the results of the averages from three trials of each experiment, with standard deviations, are presented. See Supplementary H for more experimental details.

Results. The experimental results presented in Table 1 demonstrate the efficacy of L-Reg in improving the performance of GMDG across all datasets in mDG classification tasks. Notably, more substantial improvements are observed when the GMDG baseline achieves relatively low accuracy. These observed enhancements provide empirical support for Proposition 4.1. Please see using L-Reg with basic ERM in Appendix E. For detailed insights into each domain within each dataset, please refer to Appendix H.1.

5.2 Experiments on GCD

Experimental settings. We validate our approach through training PIM additionally with L-Reg. Six image datasets are adopted to validate the feasibility of our proposed RPIM compared to other competitors, including three generic object recognition datasets, CIFAR10 [29], CIFAR100 [29] and ImageNet-100 [17]; two fine-grained datasets CUB [56] and Stanford Cars [28]; and the long-tail dataset Herbarium19 [49]. Following prior works [54, 16], we use the proposed accuracy metric from [54] of all classes, known classes, and unknown classes for evaluation. Please see a detailed description of the experimental setup in Appendix H.2.

Results. The average results across all datasets for utilizing L-Reg with PIM are presented in Table 2, while detailed dataset-specific information is available in Appendix H Table 17. The results highlight that L-Reg consistently increases the accuracy of all unknown classes across all datasets, thus confirming the validity of Proposition 4.2. However, it is notable that L-Reg may marginally compromise the performance of known classes, as it reduces the size of semantic support for deducing Y, thereby reducing the information available for known classification. Nevertheless, this compromise is deemed acceptable given the significant improvements observed for the unknown classes.

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 [38]	44.7	46.0	43.9
RankStats+ [22] (TPAMI-21)	38.6	54.6	25.6
UNO+ [19] (ICCV-21)	51.2	74.5	36.7
ORCA [13] (ICLR-22)	46.3	51.3	41.2
ORCA - ViTB16	56.7	65.6	49.9
GCD [54] (CVPR-22)	60.4	71.8	52.9
RIM [27] (NeurIPS-10)	62.0	72.5	55.4
TIM [11] (NeurIPS-20)	62.7	72.6	56.4
PIM [16] (ICCV-23)	67.4	**79.3**	59.9
PIM + L-Reg	**68.8**[1.4↑]	79.0[0.3↓]	**62.7**[2.8↑]

Table 3: 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

5.3 Experiments on mDG + GCD

Experimental settings. We utilize datasets designed for mDG tasks to conduct mDG + GCD experiments. During the training stage, only samples from seen domains are available, with half of the classes masked as unknown, and only their unlabeled data are utilized. Notably, even though all the unlabeled data originates from unknown classes during training, this prior knowledge is not assumed or constrained, aligning the setting with GCD. Similar to mDG, we adopt the leave-one-out cross-validation method. This entails testing each domain in each dataset as the unseen domain. The performance is tested on unseen domains by employing GCD metrics. To validate L-Reg's efficacy comprehensively, we re-implement four methods under the mDG + GCD setting, testing them both with and without L-Reg. The four methods include ERM, PIM, MIRO, and GMDG. ERM serves as the baseline approach without additional regularization, while PIM maximizes information without minimizing domain gaps. MIRO and GMDG focus on minimizing domain gaps, with GMDG offering a comprehensive approach in this regard. It is worth noting that PIM has been re-implemented. For further experimental details, please refer to Appendix H.3.

Results. The averaged results across all unseen domains of all datasets are summarized in Table 3. For a detailed breakdown of results for each domain in each dataset, please refer to Appendix H.3. As discussed in Proposition 4.1 and Proposition 4.2, a noticeable trend is observed wherein, as the

domain gap is gradually minimized, the improvements for unknown classes increase, with the best results achieved using GMDG with L-Reg.

L-Reg forms atomic formulas and improves interpretability. Furthermore, Fig. 5 provides visual insights into the behavior of models trained with L-Reg. Evidently, these models tend to focus on minimal semantics sufficient for class distinctions. For the known classes, the efficacy of L-Reg can be intuitively understood as extracting the minimal semantic supports for a given class label. For instance, the presence of a guitar's fingerboard, even in unseen domains, helps classify a sample as belonging to the guitar category, whose informal forms can be denoted as $h(\text{has fingerboard}, \text{is guitar}, d \in D) \to$ True and $h(\text{not has fingerboard}, \text{is guitar}, d \in D) \to$ False. For all known classes, samples with these minimal semantic supports are recognized accordingly. In contrast, if a sample lacks these minimal supports for any known class, it is very likely categorized as an unknown class. This behavior stems from Paper Eq.10 which ensures $\mathcal{A}^{y_i*} \neq \mathcal{A}^{y_j*}$ through constraining $\gamma^{y_i} \neq \gamma^{y_j}$. L-Reg further enhances the model's ability to identify minimal supports for unknown classes

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

by filtering out co-covariant features associated with other classes and thus generalizing to unseen domains. Therefore, the very interpretable features for unknown classes from unseen domains can be extracted using L-Reg. Fig. 5 (right side) demonstrates that the model with L-Reg can even extract facial features for the unknown person class and can generalize this to the unseen domain. Similarly, here we obtain (informal) atomic formulas as $h(\text{has a face}, \text{is person}, d \in D) \to$ True, $h(\text{not has a face}, \text{is person}, d \in D) \to$ False.

However, as shown in Row 3, significant domain shifts, such as those between the sketch domain and other domains, pose challenges. Specifically, the differences between the stick-figure style of sketches of persons and figures from other domains can hinder the model's ability to cluster sketches with other domains' figures when the class label is unknown. Thus, under this circumstance, the model may fail to extract meaningful features from those sketches. We acknowledge this limitation and will explore solutions in future work.

L-Reg should be applied to features from deep layers. One crucial precondition highlighted in the theoretical analysis is that L-Reg operates effectively with a representation Z, where each dimension represents independent semantics. The semantic features usually come from the deeper layers of the model

Table 4: 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**

architecture [51]. However, Table 4 shows that applying L-Reg to features from earlier layers, which may not necessarily represent semantics, leads to a degradation in performance for known classes, albeit improving performance for unknown classes. This phenomenon arises due to the potential interdependence among features from earlier layers, resulting in penalization that may hinder the capture of semantic supports essential for known classes. To ensure generalization improvements without significant compromise to the performance of known classes, we advocate for applying L-Reg specifically to features extracted from deeper layers, such as the bottleneck layer. These suggest that the compromised results observed in Table 2 could be attributed to the less depth of the model structure, which fails to provide the expected semantic features.

5.4 Apply L-Reg to congestion prediction for circuit design.

Experimental settings. We also test L-Reg in Congestion prediction on the CircuitNet [15] dataset by using CircuitFormer [63] backbone. The congestion prediction is for circuit design and benefits from logical reasoning-based approaches. All parameters, except for L-Reg, remain consistent with CircuitFormer, and we follow its metrics.
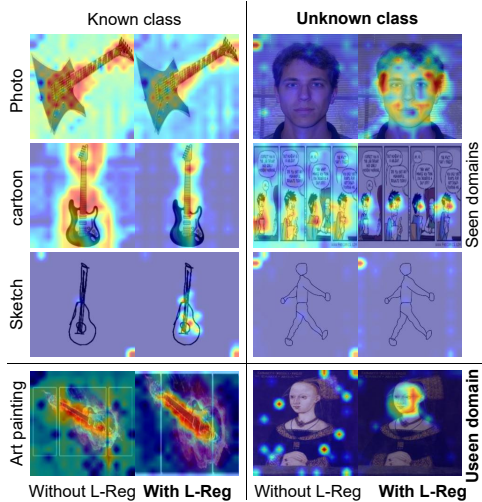
9

Results. Table 5 shows the results of prediction results on the CircuitNet dataset. We also include the results of Gpdl with UNet++ and CircuitFormer for better comparison. Notably, the improvements brought by CircuitFormer with L-Reg across all metrics, especially for the pearson metric can be observed. The consistent improvement with L-Reg across all metrics indicates L-Reg's feasibility.

Table 5: **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**

6 Related work

Logical reasoning for deep learning. Current studies focus on length generalization or symbolic reasoning in the logic-based scope. For length generalization, [1] proposes the generalization to the unseen setting, theoretically verifying that commonly used models can generalize to the unseen and degree curriculum promotes the generalization ability of the transformer, followed by [3, 2, 60]. Another branch is to improve the logical reasoning ability for abstract symbols, such as learning the logical-based temples and expecting the model to generalize to unseen samples [10, 36]. These studies are closely related to languages, such as generating longer answering sequences or solving mathematical problems in large language models, lacking explicit connections to visual tasks. [6] delves into the logical explanations in image classification by explicitly extracting logical relationships. While this logical-based approach sheds light on the interpretability of image classification models, its specific benefits for visual generalization remain relatively unexplored.

Multi-domain generalization. Current approaches for mDG in image classification focus on learning invariant representation across domains. Previous approaches like DANN [20] minimize feature divergences between source domains. CDANN [35], CIDG [34], and MDA [23] consider conditions for learning conditionally invariant features. MIRO [25] and GMDG [50] take advantage of pre-trained models to improve generalization. Specifically, in comparison to MIRO, GMDG proposes a general entropy-based learning objective for mDG and sufficiently minimizes the domain gaps, yielding better generalization results.

Generalized category discovery. Generalized category discovery, pioneered by [54], addresses unlabeled samples with both known and unknown classes. Furthermore, PIM [16] integrates InfoMax into generalized category discovery, effectively handling imbalanced datasets and surpassing GCD on both short- and long-tailed datasets.

7 Conclusion

This paper presents L-Reg, a logical regularization approach tailored for image classification tasks using logic analysis frameworks. L-Reg yields better generalization across different settings by fostering balanced feature distributions and streamlining the classification model's complexity. Rigorous theoretical analyses and empirical validations underscore its efficacy, as L-reg consistently improves generalization performance with different frameworks under various scenarios.

Limitation. L-Reg narrows the extent of semantic supports, potentially diminishing the amount of information available for classification and leading to certain trade-offs in the performance of seen datasets. This effect is evidenced by the slight decline in the accuracy of known classes when L-Reg is applied, as shown in Table 2. A similar phenomenon is observed in Fig. 5, where the model fails to recognize a person in the sketch domain lacking facial features. Analysis from Table 4 suggests that these compromises may result from improper Z. Future work should focus on mitigating potential compromises on seen datasets by exploring strategies for better capturing Z through improved model architecture design. We offer more experimental results of possible solutions to this limitation in Appendix G, such as further constraining the independence of each dimension in Z. Those results may suggest a direction for future work.

Acknowledgments

References

[1] Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In *International Conference on Machine Learning*, pages 31–60. PMLR, 2023.

[2] Emmanuel Abbe, Elisabetta Cornacchia, and Aryo Lotfi. Provable advantage of curriculum learning on parity targets with mixed inputs. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Kartik Ahuja and Amin Mansouri. On provable length and compositional generalization. *arXiv preprint arXiv:2402.04875*, 2024.

[4] Hajnal Andréka, István Németi, and Ildikó Sain. Universal algebraic logic. *Studies in Logic, Springer, due to*, 2017.

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

[6] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, and Stefano Melacci. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6046–6054, 2022.

[7] 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.

[8] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021.

[9] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

[10] Enric Boix-Adsera, Omid Saremi, Emmanuel Abbe, Samy Bengio, Etai Littwin, and Joshua Susskind. When can transformers reason with abstract symbols? *arXiv preprint arXiv:2310.09753*, 2023.

[11] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020.

[12] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021.

[13] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022.

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[15] Zhuomin Chai, Yuxiang Zhao, Wei Liu, Yibo Lin, Runsheng Wang, and Ru Huang. Circuitnet: An open-source dataset for machine learning in vlsi cad applications with improved domain-specific evaluation metric and learning strategies. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(12):5034–5047, 2023.

[16] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1729–1739, 2023.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] 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.

[19] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[21] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[22] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[23] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020.

[24] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020.

[25] Cha Junbum, Lee Kyungjae, Park Sungrae, and Chun Sanghyuk. Domain generalization by mutual-information regularization with pre-trained models. *European Conference on Computer Vision (ECCV)*, 2022.

[26] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[27] Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. *Advances in neural information processing systems*, 23, 2010.

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[30] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[33] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[34] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[35] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018.

[36] Zenan Li, Yunpeng Huang, Zhaoyu Li, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, and Jian Lu. Neuro-symbolic learning yielding logical constraints. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Ziyue Li, Kan Ren, Xinyang Jiang, Yifei Shen, Haipeng Zhang, and Dongsheng Li. Simple: Specialized model-sample matching for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2022.

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

[39] Ofir Moshe, Gil Fidel, Ron Bitton, and Asaf Shabtai. Improving interpretability via regularization of neural activation sensitivity. *arXiv preprint arXiv:2211.08686*, 2022.

[40] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[41] Konstantinos Panagiotis Panousis, Dino Ienco, and Diego Marcos. Sparse linear concept discovery models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2767–2771, 2023.

[42] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[43] Wilhelm Rödder. Conditional logic and the principle of entropy. *Artificial Intelligence*, 117(1):83–106, 2000.

[44] 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. *arXiv preprint arXiv:1911.08731*, 2019.

[45] 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.

[46] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[47] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

[48] 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*, pages 443–450. Springer, 2016.

[49] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019.

[50] 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)*, pages 23512–23522, June 2024.

[51] 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*, volume 38, pages 5098–5107, 2024.

[52] Irene Tsapara and György Turán. Learning atomic formulas with prescribed properties. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 166–174, 1998.

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

[54] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.

[55] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[56] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[57] Chunyang Wu, Mark JF Gales, Anton Ragni, Penny Karanasou, and Khe Chai Sim. Improving interpretability and regularization in deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):256–265, 2017.

[58] Mike Wu, Michael Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[59] Mike Wu, Sonali Parbhoo, Michael Hughes, Ryan Kindle, Leo Celi, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Regional tree regularization for interpretability in deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6413–6421, 2020.

[60] Changnan Xiao and Bing Liu. A theory for length generalization in learning to reason. *arXiv preprint arXiv:2404.00560*, 2024.

[61] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[62] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[63] Jialv Zou, Xinggang Wang, Jiahao Guo, Wenyu Liu, Qian Zhang, and Chang Huang. Circuit as set of points. *Advances in Neural Information Processing Systems*, 36, 2024.

A Broader impact

Our regularization term based on logic for image classification offers significant potential beyond academia. By integrating logical constraints, our approach enhances model robustness, interpretability, and ethical alignment. This translates into improved performance on real-world tasks such as disease diagnosis in healthcare and mitigating biases in decision-making systems. Our work fosters interdisciplinary collaboration and contributes to the responsible deployment of AI technologies, ultimately benefiting society through enhanced efficiency, fairness, and transparency in machine learning applications.

B Details of the logical framework for visual classification task

We provide more details of the connections between logical reasoning and visual classification tasks.

Definition B.1. Following [4], a logic $\mathcal{L}$ is a five-tuple defined in the form:

$$\mathcal{L} = \langle F_{\mathcal{L}}, M_{\mathcal{L}}, \models_{\mathcal{L}}, mng_{\mathcal{L}}, \vdash_{\mathcal{L}} \rangle, \tag{16}$$

where

- $F_{\mathcal{L}}$ is a set of all formulas of $\mathcal{L}$. $F_{\mathcal{L}}$ arbitrarily refers to any collections that can be 'expressed' by language $\mathcal{L}$. Therefore, $F_{\mathcal{L}}$ could be not only a collection of languages but also images and labels (X, Y) for computer vision cases.

- $M_{\mathcal{L}}$ is a class called the class of all models (or possible worlds) of $\mathcal{L}$; intuitively, this can be considered as different domains D of X.

- $\models_{\mathcal{L}}$ is a binary relation, $\models_{\mathcal{L}} \subseteq M_{\mathcal{L}} \times F_{\mathcal{L}}$, called the validity relation of $\mathcal{L}$. For example, 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 $\mathcal{L}$: Intuitively, $mng_{\mathcal{L}}$ extracts the meaning of the expressions can be understood as the 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 $mng_{\mathcal{L}}$, such as the estimation criteria in the machine learning system.

Accordingly and still following [4], a good general logic is defined as:

Definition B.2 (General logic)**.** : A general logic is a class:

$$\mathcal{L}^* := \left\langle \mathcal{L}^P : P \in Sig \right\rangle, \tag{17}$$

where Sig is a class of sets; $\mathcal{L}^P = \left\langle F_{\mathcal{L}}^P, M_{\mathcal{L}}^P, \models_{\mathcal{L}}^P, mng_{\mathcal{L}}^P, \vdash_{\mathcal{L}}^P, \right\rangle$ is a compositional logic in the sense of Definition B.1 for $P \in Sig$, and for any sets $P, Q \in Sig$ satisfies the following conditions:

1. P is the set of atomic formulas of $\mathcal{L}^P$.

2. $Cn(\mathcal{L}^P) = Cn(\mathcal{L}^Q) := Cn(\mathcal{L}^*)$ where $Cn(\cdot)$ is called the set of logical connectives of the given logic (these are operation symbols with finite or infinite ranks).

3. Any bijection $f : P ß Q$ that extends to a bijection between the tautological formula algebras of $\mathcal{L}^P$ and $\mathcal{L}^Q$ induces an isomorphism between $\mathcal{L}^P$ and $\mathcal{L}^Q$.

4. If $P \subseteq Q$, then $\mathcal{L}^P$ is a sublogic of $\mathcal{L}^Q$.

5. For any $P \in Sig$ and set H, there is a $P' \in Sig$ such that P' is disjoint from H and $\mathcal{L}^{P'}$ is an isomorphic copy of $\mathcal{L}^P$.

6. The union of a system $P_i, i \in I$ of pairwise disjoint sets P_i from Sig belongs to Sig, whenever I is not empty. Let $\mathfrak{Fr}(\cdot)$ denotes free algebra, $\mathrm{Alg}_m(\mathcal{L})$ represents $\{mng_{\mathfrak{M}}(\mathfrak{F}) : \mathfrak{M} \in M\}$ where $\mathfrak{F}$ denotes the term algebra. Further, the tautological congruence of the logic belonging to the disjoint union P is generated in $\mathfrak{Fr}\left(\mathrm{Alg}_m\left(\mathcal{L}^P\right), P\right)$ as a

congruence by the union of the tautological congruence relations of the logics belonging to $P_i, i \in I$.

7. Sig contains at least one non-empty set.

Our L-Reg aims to regularize the semantics extracted by g and the classifier to satisfy condition 1.

$\vdash_{(h \circ g(X), Y)} = \models_{(g(X_s), Y_s)}$ **in Eq. (6) can be safely omitted in the rest of the paper.** Consider the logic formed on X, Y: $\mathcal{L}_{(X_s, Y_s)} = \langle F_{(X_s, Y_s)}, D, \models_{(X_s, Y_s)}, h, \vdash_{(h(X), Y)} \rangle$. Assume we want to study the logic of $\vdash$ which can be defined in the form of $\mathcal{L}_\vdash \overset{\text{def}}{:=} \langle F_{X_s, Y_s}, D_\vdash, h_\vdash, \models_\vdash \rangle$, where $D_\vdash, h_\vdash, \models_\vdash$ are pseudo-components associated with $\vdash$. Particularly, $D_\vdash$ is a subset of all possible world/domains from $F_{(X_s, Y_s)}$: $D_\vdash \overset{\text{def}}{=} \{T \subseteq F_{(X_s, Y_s)} : T \text{ is closed under } \vdash_{(h(X), Y)} \}$. For any $T \in D_\vdash$ and $a \in F_{(X_s, Y_s)}$, it has $h_\vdash(a, T) \overset{\text{def}}{=} \{b \in F : T \vdash (a \leftrightarrow b)\}$. Further, $\models_\vdash$ in $T \in D_\vdash$ is defined as $T \models_\vdash a \overset{\text{def}}{\Leftrightarrow} a \in T$. [4] points out that the following condition is almost always satisfied: (Cond) $\forall a, b \in F_\vdash, d \in D_\vdash$, we have $(h_\vdash(a, d) = h_\vdash(b, d))$ and $d \models_\vdash a \Rightarrow d \models_\vdash b$. Therefore, the semantical consequence relation induced by $\models_\vdash$ coincides with the original syntactical $\vdash_{(h \circ g(X), Y)}$ while Cond holds. Due to that $D_\vdash \subseteq D$, it infers that $\models_{(g(X_s), Y_s)}$ coincides with $\models_\vdash$. Therefore, $\vdash_{(h \circ g(X), Y)} = \models_{(g(X_s), Y_s)}$ can be safely omitted in the rest of the paper.

C Details of proofs

Proposition C.1 (L-Reg reduces the complexity of the model, promoting data-shift generalization performance.)**.** *Assume the domain gap is well minimized. Consider a f^* is the target model that generalizes to the unseen with the lowest complexity. There are $f^R_{(X_s, Y_s)}, f_{(X_s, Y_s)}$ trained under the setting of data-shift generalization (i.e., (X_s, Y_s) is accessible and $\mathcal{Y}_s = \mathcal{Y}_u$), it has that:*

$$GL(f^R_{(X_s, Y_s)}, f^*, X_u) \leq GL(f_{(X_s, Y_s)}, f^*, X_u), \tag{18}$$

Proof. We assume the loss is achieved for the tractable form by minimizing the mean squared error. In that case, we have $f^*_{(X_s, Y_s)}$ for the given training set as:

$$f^*_{(X, Y)} = (g(X)^T g(X))^{-1} h \circ g(X)^T Y_s = (Z^T Z)^{-1} h(Z^T) Y, \tag{19}$$

In comparison to f^*, $f_{(X_s, Y_s)}$ for the given seen sets is as:

$$f_{(X_s, Y_s)} = (Z_s^T Z_s)^{-1} h(Z_s^T) Y_s, \tag{20}$$

and $f^R_{(X_s, Y_s)}$ is derived from $f_{(X_s, Y_s)}$, where Z_s is constrained additionally by L-Reg and the constrained Z_s is denoted as Z_s^R:

$$f^R_{(X_s, Y_s)} = (Z_s^{R\,T} Z_s)^{-1} h(Z_s^{R\,T}) Y_s. \tag{21}$$

For simplification, we denote $(Z^T Z)^{-1} Z^T$, $(Z_s^T Z_s)^{-1} Z_s^T$, and $(Z_s^{R\,T} Z_s)^{-1} Z_s^{R\,T}$ as $\mathcal{N}^*, \mathcal{N}$, and $\mathcal{N}^R$, respectively.

The form of $\mathcal{N}$. For multi-domain generalization, the model is tested on the unseen domain, referring that X_u contains some unseen semantics besides the seen: $Z_s \sim \mathcal{Z}_s, Z_u \sim \mathcal{Z}_u, \mathcal{Z}_s \neq \mathcal{Z}_u, \mathcal{Z}_s \cap \mathcal{Z}_u \neq \emptyset$. Considering each dimension of Z represents a specific semantics, we denote Γ as the dimensions of Z that contain the seen semantics support in X_s and $\bar{\Gamma}$ for the unseen, we can decompose $\mathcal{N}$ as:

$$\mathcal{N} = \begin{bmatrix} \Gamma^T \bar{\Gamma} & \Gamma^T \bar{\Gamma} \\ \bar{\Gamma}^T \Gamma & \bar{\Gamma}^T \bar{\Gamma} \end{bmatrix}^{-1} [h(\Gamma)\ h(\bar{\Gamma})]^T. \tag{22}$$

The form of $\mathcal{N}^*$. Assume Γ already contains semantic support for deducting Y; thus, $\bar{\Gamma}$ would not affect the deduction of Y. In such case, it has that $\Gamma^T \bar{\Gamma} = \mathbf{0}$ and $\bar{\Gamma}^T \Gamma = \mathbf{0}$ and $\bar{\Gamma}^T \bar{\Gamma} = \mathbf{1}$ where $\mathbf{0}, \mathbf{1}$

denote zero matrix and identity matrix:

$$
\begin{aligned}
\mathcal{N}^* &= \begin{bmatrix} \Gamma^T\Gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}^{-1} [h(\Gamma)\ h(\bar{\Gamma})]^T \\
&= \begin{bmatrix} (\Gamma^T\Gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}^{-1} [h(\Gamma)\ h(\bar{\Gamma})]^T \\
&= \begin{bmatrix} (\Gamma^T\Gamma)^{-1}h(\Gamma) \\ h(\bar{\Gamma}) \end{bmatrix},
\end{aligned}
\tag{23}
$$

where we also expect $h(\bar{\Gamma}) = \mathbf{0}$ so that z_u does not influence the deduction. We now have $\mathcal{N}^*$:

$$
\mathcal{N}^* = \begin{bmatrix} \Gamma^T\Gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}^{-1} [h(\Gamma)\ h(\bar{\Gamma})]^T, \ \text{s.t.,}\ h(\bar{\Gamma}) = \mathbf{0}.
\tag{24}
$$

Note that for $\mathcal{N}$ in $f_{(X_s, Y_s)}$, $\Gamma^T\bar{\Gamma}$ and $\bar{\Gamma}^T\Gamma$ are not constrained. Please refer to Lemma C.2. Furthermore, $h(\bar{\Gamma})$ is also not constrained.

The form of $\mathcal{N}^R$. Now we discuss the trainable $\mathcal{N}^R$ obtained with the application of L-Reg. The form of $\mathcal{N}^R$ is similar to $\mathcal{N}^*$. However, Eq. (11) indicates that L-Reg minimizes $||\Gamma^T\bar{\Gamma}||_2$ and $||\bar{\Gamma}^T\Gamma||_2$ through $-H(Y|g(\bar{\Gamma})), D)$ and also minimizing $||h(\bar{\Gamma})||_2$:

$$
\mathcal{N}^R = \begin{bmatrix} \Gamma^T\bar{\Gamma} & \Gamma^T\bar{\Gamma} \\ \bar{\Gamma}^T\Gamma & \bar{\Gamma}^T\bar{\Gamma} \end{bmatrix}^{-1} [h(\Gamma)\ h(\bar{\Gamma})]^T, \ \text{s.t.,}\ \min ||\Gamma^T\bar{\Gamma}||_2 + ||\bar{\Gamma}^T\Gamma||_2 + |h(\bar{\Gamma})||_2.
\tag{25}
$$

Compare $GL(f^R_{(X_s, Y_s)}, f^*, X_u)$ **with** $GL(f_{(X_s, Y_s)}, f^*, X_u)$**.** By comparing the forms of $\mathcal{N}^R, \mathcal{N}$ and $\mathcal{N}^*$, it is obvious that $||\mathcal{N}^R - \mathcal{N}^*||_2 \leq ||\mathcal{N} - \mathcal{N}^*||_2$. Therefore, we have that: $GL(f^R_{(X_s, Y_s)}, f^*, X_u) \leq GL(f_{(X_s, Y_s)}, f^*, X_u)$. □

Lemma C.2 (Minimizing $H(Y|g(X), D) + \mathcal{R}$ solely may cause generalization degradation)**.** *Minimizing $H(Y|g(X), D) + \mathcal{R}$ solely without L-Reg may conflict with $\max_{h,g} H(Y|g(\bar{\Gamma}), D)$, causing invalid semantics for decision process and degrading the generalization.*

Proof. We have the following relationship for $H(Y|g(z), D)$:

$$
\begin{aligned}
H(Y|g(z), D) &= H(Y|g(\bar{\Gamma}), g(\Gamma), D) \\
H(Y, g(\Gamma)|g(\bar{\Gamma}), D) - H(g(\Gamma)|g(\bar{\Gamma}), D) &= H(Y|g(\bar{\Gamma}), g(\Gamma), D) + H(g(\bar{\Gamma})|g(\Gamma), D).
\end{aligned}
\tag{26}
$$

Since the independence between $\{z_i\}_{i=1}^M$ is unconstrained, $H(Y, g(\Gamma)|g(\bar{\Gamma}), D)$ may cause that Y can be deduced from $\bar{\Gamma}$. Therefore, $\Gamma^T\bar{\Gamma}$ and $\bar{\Gamma}^T\Gamma$ are not constrained even when the domain gap is minimized where $|D| = 1$, causing the sub-optimal generalization. □

D One toy example

We present a simplified informal illustrative example to compare the efficacy of our proposed L-Reg against conventional L1 and L2 regularization methods. As depicted in Fig. 6, the ground truth (GT) image represents the underlying data, generated according to $f^*(x_1, x_2) = \sin(2\pi x_1) \cdot \sin(2\pi x_2)$, where x_1 and x_2 denote the horizontal and vertical coordinates respectively, and the pixel color corresponds to the value of $f^*(x_1, x_2)$. The training domain is delineated by the black box, while the testing domain encompasses the area outside of this boundary.

For our experiments, we use a 6-linear-layer size-110 ReLU model network. Mean squared error serves as the loss function.

Our experimental results reveal that L-Reg enhances the model's ability to extrapolate beyond the training domain. Notably, our proposed L-Reg demonstrates superior extrapolative capabilities compared to traditional L_1 and L_2 regularization methods. This observation highlights the efficacy of L-Reg in fostering improved generalization.
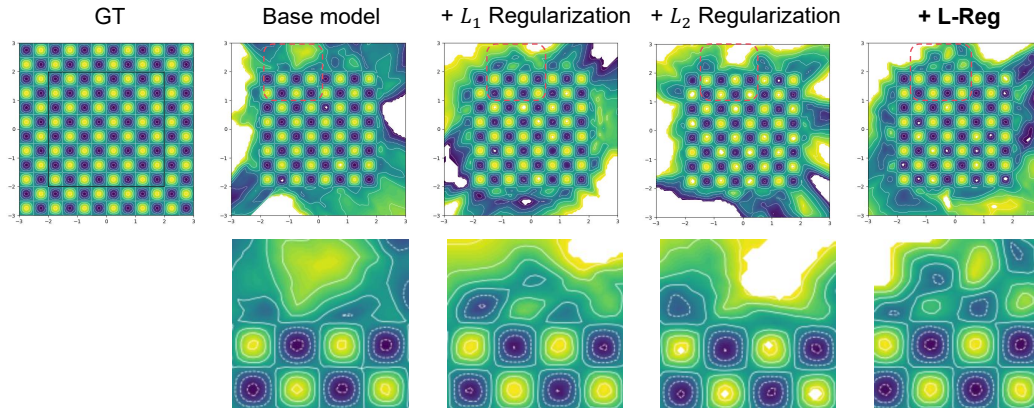
Figure 6: Prediction visualizations of MLP with different regularization terms.

E Apply L-Reg to ERM Baseline for mDG

To further validate L-Reg's efficacy for mDG, we use ERM as the baseline on the TerraIncognita dataset. For a fair comparison, all experiments share the same hyperparameter settings and use the Regnety-16gf backbone. Original ERM results are also included alongside our reproduced results. The results in Table 8 reveal that ERM with L-Reg significantly improves mDG performance (from 49.9% to 52.9%).

F Compare L-Reg with more regularization terms

We also compare L-Reg with other regularization terms: The Ortho-Reg - the orthogonality regularization that constrains the independence of each dimension of the semantic feature z; and Sparsity - implemented as Bernoulli Sample of the latent features from the sparse linear concept discovery models [41] on our used PIM backbone. To investigate this fairly, we re-implemented the Bernoulli Sample of the latent features from the Sparse Linear Concept Discovery Models [41] on the same PIM backbone that we used, to achieve the sparsity. Table 6&Table 7 demonstrate that L-Reg outperforms Ortho-Reg and Sparsity.

Especially, while a common sparse concept model may be able to achieve $\gamma^y \psi = z^y$ by filtering irrelevant features through the sparsity, it may not ensure $\gamma^{y_i} \neq \gamma^{y_j}$, which is crucial for disentangling features used for predicting different classes. This limitation can potentially lead to degradation in generalization performance for common sparse concept models. 6&7 indicate that while L-Reg consistently achieves overall improvement, the sparse concept-based approach does not consistently improve generalization, validating the aforementioned difference.

G Limitation of L-Reg and possible solutions

As analyzed and discussed in the paper, L-Reg is based on the precondition that each dimension of the latent features represents an independent semantic.

We hypothesize this is due to the fact that our L-Reg is derived based on the precondition that $z^i, z^j \in z, I \neq j$ is independent of each other. This condition holds for most deep-layer features but may not apply to shallow layers. Thus, applying L-Reg to the semantic features from the deep layers may improve the performance for unknown classes without negatively impacting known classes.

Derived from this hypothesis, another possible solution is further regularizing the independence, which may lead to further improvements. To validate this hypothesis, we test L-Reg by reinforcing independence with Ortho-Reg. MDG results in Table 8 and GCD results in Table 6&Table 7 show that combining L-Reg with Ortho-Reg leads to further improvements, whereas Ortho-Reg alone may not guarantee improvements. These findings support our hypothesis and suggest that L-Reg, particularly

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

when applied to deep layers or in conjunction with Ortho-Reg, is beneficial. This suggests a direction for future work.

H More experimental details and results

All experiments can be conducted on one NVIDIA GeForce RTX 3090 GPU.

H.1 Multi-domain generalization

Competitors. We listed results from previous important work in the mDG field for better validation. They are: MMD [33], Mixstyle [62], GroupDRO [44], IRM [5], ARM [61], VREx [30], CDANN [35], DANN [20], RSC [24], MTL [8], MLDG [31], Fish [46], ERM [53], SagNet [40], SelfReg [26], CORAL [48], mDSDI [12], MIRO [25], and GMDG [50]. Among them, GMDG is treated as our baseline since it sufficiently minimizes the domain gaps.

Datasets. We use PACS (4 domains, 9,991 samples, 7 classes) [32], VLCS (4 domains, 10,729 samples, 5 classes) [18], OfficeHome (4 domains, 15,588 samples, 65 classes) [55], TerraIncognita

Table 8: **Results of mDG:** Results of using ERM as the baseline. We use the ERM method as the baseline to test L-Reg's efficacy. Ortho-Reg: orthogonality regularization. This table includes results: (1) The improved performance of L-Reg on ERM baseline. (2) Comparison between L-Reg and the Ortho-Reg on ERM baseline. (3) Using L-Reg and Ortho-Reg together yields further promotion, validating our 'improper z' hypothesis in the Paper limitation part. The used dataset is TerraIncognita. All experiments share the same hyperparameters except the added regularization term. Each group of experiments is run with seeds [0,1,2], and the averaged results for each domain and additionally with the standard deviation (Std) are reported.

TerraIncognita	Location 100	Location 38	Location 43	Location 46	Avg ± Std.
ERM	54.3	42.5	55.6	38.8	47.8
ERM Reproduced	50.6	49.7	58	41.2	49.9±3.6
+Ortho-Reg	50.7	52.6	60.5	42.7	51.6±2.5
+L-Reg	52.7	51.7	61.3	45.8	52.9±4.2
+L-Reg+Ortho-Reg	61.5	48.6	60.3	44	53.6±0.5

Table 9: Parameters for mDG task

Use RegNetY-16GF	lr mult	α
TerraIncognita	2.5	1e-3
OfficeHome	0.1	1e-3
VLCS	0.1	1e-4
PACS	0.1	5e-4
DomainNet	5.0	1e-3

(TerraIncognita, 4 domains, $24,778$ samples, 10 classes) [7], and DomainNet (6 domains, 586,575 samples, 345 classes) [42].

Training details. We use GMDG [50] as our baseline. Especially, we use all loss terms proposed in GMDG as L_{main}. The training procedure is the same as MIRO [25] and GMDG. We use seeds $0, 1, 2$ for all three trails training.

Parameters. We adhere to the parameters proposed by GMDG, particularly focusing on its recommended loss terms. Furthermore, we provide a detailed listing of the hyper-parameters pertaining to L-Reg, along with the tuned 'lr mult', as outlined in Table 9, to facilitate the reproducibility of our results.

Evaluation metric. The models undergo training on known domains and subsequent testing on unseen domains. For each trial, a distinct domain within the datasets is designated as the unseen domain. The evaluation metric reports the prediction accuracy achieved on these unseen domains. The aggregated results across all unseen domains within the datasets provide a comprehensive assessment of the algorithm's performance in domain generalization for the given datasets.

More results. Results of each domain for each dataset are presented in Tables 10 to 14.

H.2 Generalized category discovery

Competitors. We compare our proposed method with existing generalized category discovery methods: GCD [54], and PIM [16]. In particular, PIM based on information maximization is the current state-of-the-art (SOTA) generalized category discovery method. Additionally, the traditional machine learning method, k-means [38]; three novel category discovery methods: RankStats+ [22], UNO+ [19], ORCA [13]; and several information maximization methods: RIM [27], and TIM [11] are adapted for generalized category discovery as competitors. The results of the modified novel category discovery methods are reported in [54], and the modified information maximization methods are reported in [16].

Usage details of datasets for GCD. Following the protocols of GCD and PIM [54, 16], the initial training set of each dataset is divided into labeled and unlabeled subsets; samples from half of the classes are assigned as unlabeled, and their labels are not used for training. Specifically, half of the image samples from known classes are allocated to the labeled subset, while the remaining half are assigned to the unlabeled subset. Additionally, the unlabeled subset includes all image samples from

the novel classes in the original dataset. As a result, the unlabeled subset consists of instances from K different classes. The detailed statistics of datasets are listed in Table 15.

Training details. Consistent with PIM, we utilize latent features extracted by the feature encoder DINO (VIT-B/16) [14] that is pre-trained on ImageNet [17] through self-supervised learning. The losses proposed in PIM are treated as L_{main}. The original PIM freezes the feature extractor during the training, directly using the pre-saved extracted features as the model input. For a fair comparison, we only added one linear layer as g on the extracted features, which is the minimal modification.

L_2 **(weight decay) value searching.** For a more fair comparison, we conduct weight decay value searching to ensure that the weight of L_2 is the best. To address this, we devised a methodology for weight decay searching involving the construction of smaller labeled and unlabeled subsets derived solely from the labeled data. To conduct parameter searching, we split the labeled samples to construct a 'smaller' sub-labeled and sub-unlabeled set. Specifically, we take 50% of the samples from known classes as sub-unlabeled samples from unknown classes. Additionally, we take 25% of the samples from the remaining 50% of known classes as sub-unlabeled samples from known classes. The remaining samples are treated as sub-labeled samples. Hyper-parameters are then searched on these sub-labeled and sub-unlabeled sets.

Parameters of L-Reg. The hyper-parameters of L-Reg values are shown in Table 16.

Evaluation metric. Following prior works [54, 16], we use the proposed accuracy metric from [54] of all classes, known classes, and unknown classes for evaluation.

More results. The results for each dataset are presented in Table 17. It is evident that L-Reg yields enhanced performance across half of the datasets for both known and unknown classes. On the remaining datasets, while L-Reg may slightly compromise the performance of known classes, it demonstrates significant improvements in the unknown classes, resulting in an overall enhancement in the performance across all classes.

More ablation results. Due to the introduction of tuned weight decay and the additional g component, we have conducted ablation studies to assess their impact. The results are summarized in Table 18. It is observed that the baseline model utilizing the tuned weight decays performs slightly better than the original weight decay settings. Notably, the tuned weight decays contribute to improvements in unknown classes while often leading to slight decreases in known classes across most datasets. Inclusion of the proposed extra component g results in marginal improvements in both known and unknown classes compared to the tuned baseline. Our proposed L-Reg demonstrates significant improvements specifically in the unknown classes, thereby corroborating Proposition 4.2. However, as discussed in the main paper, it is acknowledged that L-Reg may entail compromises in the performance of known classes.

H.3 Combination of multi-domain generalization and generalized category discovery

Datasets. We leverage the datasets utilized in mDG tasks to construct the mDG+GCD datasets. Specifically, during the seen domains of training, labels from approximately half of the classes are masked. For instance, in the PCAS dataset comprising 7 classes, classes labeled within the range $[0, 1, 2, 3]$ are retained, while classes in $[4, 5, 6]$ are masked. It is noteworthy that data categorized as unknown classes in our setup are from unknown classes. However, we acknowledge that this prior is not explicitly known. **To align with the GCD setting, we operate under the assumption that the unlabeled set may potentially include samples from known classes.** Consequently, we refrain from constraining the model by mandating that unlabeled data be classified solely as unknown classes. This adjustment introduces a more challenging generalization scenario.

Training details. For all experiments, the implementation directly adds L-Reg to their previously proposed loss sets. The models are trained with the aforementioned labeled and unlabeled sets from the seen domains and tested on the samples from the unseen domain.

Parameters. We include all the parameters for reproducing our experiments in the code. Please refer to the code for details.

Evaluation metric. We use the same metric from the GCD task for the mDG+GCD task. Similarly, the metrics include the accuracy for all, known and unknown classes.

More results. The averaged results of each dataset are exhibited in Table 19, while the detailed results of each dataset are presented in Tables 20 to 24.

H.4 More GradCAM visualizations

We provide more visualized examples of L-Reg. Examples of known classes can be seen in Figs. 7 to 10 and unknown classes in Figs. 11 and 12. Compromises in known sets, as discussed in the limitations, can be seen in Figs. 8 and 12.

Table 10: MDG experiments on TerraIncognita: More results of full GMDG+L-Reg for each category.

TerraIncognita	Location 100	Location 38	Location 43	Location 46	Avg.
ERM [21]	54.3	42.5	55.6	38.8	47.8
MIRO [25] (use ResNet-50)	-	-	-	-	50.4
GMDG [50] (use ResNet-50)	59.8±1.0	45.3±1.7	57.1±1.8	38.2±5	50.1±1.2
MIRO [25] (use RegNetY-16GF)	-	-	-	-	58.9±1.3
GMDG [50] (use RegNetY-16GF)	73.3±3.3	54.7±1.4	67.1±0.3	48.6±6.5	60.7±1.8
GMDG + **L-Reg** (use RegNetY-16GF)	**73.9**±0.8	**57.1**±2.3	**67.9**±1.1	**52.7**±4.0	**62.9**±0.9

Table 11: MDG experiments on OfficeHome: More results of full GMDG+L-Reg for each category.

OfficeHome	art	clipart	product	real	Avg.
ERM [21]	63.1	51.9	77.2	78.1	67.6
MIRO [25] (use ResNet-50)	-	-	-	-	70.5±0.4
GMDG [50] (use ResNet-50)	68.9±0.3	56.2±1.7	79.9±0.6	82.0±0.4	70.7±0.2
MIRO [25] (use RegNetY-16GF)	-	-	-	-	80.4±0.2
GMDG [50] (use RegNetY-16GF)	**79.7**±1.6	67.7±1.8	87.8±0.8	87.9±0.7	80.8±0.6
GMDG + **L-Reg** (use RegNetY-16GF)	78.4±0.3	**69.3**±0.7	**87.9**±0.6	**88.0**±0.8	**80.9**±0.5

Table 12: MDG experiments on VLCS: More results of full GMDG+L-Reg for each category.

VLCS	caltech101	labelme	sun09	voc2007	Avg.
ERM [21]	97.7	64.3	73.4	74.6	77.3
MIRO [25] (use ResNet-50)	-	-	-	-	79.0±0.0
GMDG [50] (use ResNet-50)	98.3±0.4	65.9±1	73.4±0.8	79.3±1.3	79.2±0.3
MIRO [25] (use RegNetY-16GF)	-	-	-	-	79.9±0.6
GMDG [50] (use RegNetY-16GF)	97.9±1.3	66.8±2.1	**80.8**±1	**83.9**±1.8	82.4±0.6
GMDG + **L-Reg** (use RegNetY-16GF)	**98.6**±0.1	**67.1**±0.1	80.7±0.7	83.0±0.8	**82.4**±0.0

Table 13: MDG experiments on PACS: More results of full GMDG+L-Reg for each category.

PACS	art_painting	cartoon	photo	sketch	Avg.
ERM [21]	84.7	80.8	97.2	79.3	84.2
MIRO [25] (use ResNet-50)	-	-	-	-	85.4±0.4
GMDG [50] (use ResNet-50)	84.7±1.0	81.7±2.4	97.5±0.4	80.5±1.8	85.6±0.3
MIRO [25] (use RegNetY-16GF)	-	-	-	-	97.4±0.2
GMDG [50] (use RegNetY-16GF)	97.5±1.0	97.0±0.2	**99.4**±0.2	95.2±0.4	97.3±0.1
GMDG + **L-Reg** (use RegNetY-16GF)	**97.6**±0.8	**97.1**±0.3	99.3±0.2	**95.3**±0.9	**97.4**±0.2

Table 14: MDG experiments on DomainNet: More results of full GMDG+L-Reg for each category.

DomainNet	clipart	info	painting	quickdraw	real	sketch	Avg.
ERM [21]	50.1	63.0	21.2	63.7	13.9	52.9	44.0
MIRO [25] (use ResNet-50)	-	-	-	-	-	-	44.3±0.2
GMDG (use ResNet-50)	63.4±0.3	22.4±0.4	51.4±0.4	13.4±0.8	64.4±0.3	52.4±0.4	44.6±0.1
MIRO [25] (use RegNetY-16GF)	-	-	-	-	-	-	53.8±0.1
GMDG (use RegNetY-16GF)	74.0±0.3	39.5±1.5	61.5±0.3	**16.3**±1.2	73.9±1.5	62.8±2.4	54.6±0.1
GMDG + **L-Reg** (use RegNetY-16GF)	**74.1**±0.1	**42.6**±1.0	**62.3**±2.9	12.7±0.9	**75.9**±0.8	**64.6**±0.2	**55.4**±0.0

Table 15: Statistics of datasets.

	CUB	Standford Cars	Herbarium19	CIFAR10	CIFAR100	ImageNet-100
Known classes	100	98	341	5	80	50
Seen data	1.5K	2.0K	8.9K	12.5K	20K	31.9K
Known classes	200	196	683	10	100	100
Unseen data	4.5K	6.1K	25.4K	37.5K	30K	95.3K

Table 16: Tuned weight decay values for each dataset.

	CUB	Standford Cars	Herbarium19	CIFAR10	CIFAR100	ImageNet-100
Tuned weighted decay	0.02/2	0.02/2	0.02/2	0.05/2	0.005/2	0.005/2
α of L-Reg	0.1	0.001	0.2	0.01	0.0025	0.01

Table 17: GCD results: Accuracy scores across fine-grained and generic PIM datasets with our L-Reg and other competitors. The best results of each group are highlighted in **bold**. Improvement and degradation in our approach from PIM are highlighted in red and blue, respectively.

	CUB			Stanford Cars			Herbarium19		
Approach	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
K-means	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+ [22] (TPAMI-21)	33.3	51.6	24.2	28.3	61.8	12.1	27.9	55.8	12.8
UNO+ [19] (ICCV-21)	35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	**53.7**	14.7
ORCA [13] (ICLR-22)	27.5	20.1	31.1	15.9	17.1	15.3	22.9	25.9	21.3
ORCA [13] - ViTB16	38.0	45.6	31.8	33.8	52.5	25.1	25.0	30.6	19.8
GCD [54] (CVPR-22)	**51.3**	**56.6**	**48.7**	**39.0**	57.6	**29.9**	**35.4**	51.0	**27.0**
InfoMax based methods									
RIM [27] (NeurIPS-10)	52.3	51.8	52.5	38.9	57.3	30.1	40.1	**57.6**	30.7
TIM [11] (NeurIPS-20)	53.4	51.8	54.2	39.3	56.8	30.8	40.1	57.4	30.7
PIM [16] (ICCV-23)	62.7	**75.7**	56.2	43.1	**66.9**	31.6	42.3	56.1	34.8
PIM + L-Reg (Ours)	**65.3**$^{2.6\uparrow}$	76.0 $^{0.3\uparrow}$	**60.0**$^{3.8\uparrow}$	**44.8**$^{1.7\uparrow}$	66.0$^{1.4\downarrow}$	**34.6**$^{3.0\uparrow}$	**43.7**$^{2.4\uparrow}$	55.8$^{0.3\downarrow}$	**37.2**$^{1.6\uparrow}$

	CIFAR10			CIFAR100			ImageNet-100		
Approach	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
K-means	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ [22] (TPAMI-21)	46.8	19.2	60.5	58.2	**77.6**	19.3	37.1	61.6	24.8
UNO+ [19] (ICCV-21)	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	**95.0**	57.9
ORCA [13] (ICLR-22)	88.9	88.2	89.2	55.1	65.5	34.4	67.6	90.9	56.0
ORCA [13] - ViTB16	**97.1**	96.2	**97.6**	69.6	76.4	56.1	**76.5**	92.2	**68.9**
GCD [54] (CVPR-22)	91.5	97.9	88.2	**70.8**	77.6	**57.0**	74.1	89.8	66.3
InfoMax based methods									
RIM [27] (NeurIPS-10)	92.4	**98.1**	89.5	73.8	78.9	63.4	74.4	91.2	66.0
TIM [11] (NeurIPS-20)	93.1	98.0	90.6	73.4	78.3	63.4	76.7	93.1	68.4
PIM [16] (ICCV-23)	94.7	97.4	93.3	78.3	84.2	66.5	83.1	**95.3**	77.0
PIM + L-Reg(Ours)	**94.8**$^{0.1\uparrow}$	**97.6** $^{0.2\uparrow}$	**93.4**$^{0.1\uparrow}$	**80.8**$^{2.5\uparrow}$	**84.6**$^{0.2\uparrow}$	**73.2**$^{6.7\uparrow}$	**83.4**$^{0.3\uparrow}$	94.0$^{1.3\downarrow}$	**78.0**$^{1.0\uparrow}$

Table 18: GCD results: Accuracy scores across fine-grained and generic datasets of each setting. The best results are highlighted in **bold**. To eliminate the impact of hyper-parameters on performance, we also present the results of PIM with tuned hyper-parameters (termed baseline tuned). L_{main} denotes the losses used in PIM. g denotes the transformation applied to the input features.

		CUB			Stanford Cars			Herbarium19		
ID	Settings	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
1	Baseline (L_{main})	62.7	75.7	56.2	43.1	66.9	31.6	42.3	56.1	34.8
2	Baseline tuned (L_{main})	64.8	75.1	59.6	42.6	59.3	34.6	43.1	57.6	35.4
6	$L_{main} + g$	64.9	**76.7**	58.9	44.7	65.8	34.6	43.0	**57.4**	35.2
9	**Ours** ($L_{main}+h+L_{L-Reg}$)	**65.3**	76.0	**60.0**	**44.8**	66.0	34.6	**43.7**	55.8	**37.2**

		CIFAR10			CIFAR100			ImageNet-100		
ID	Settings	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
1	Baseline (L_{main})	94.7	97.4	93.3	78.3	84.2	66.5	83.1	**95.3**	77.0
2	Baseline tuned (L_{main})	95.0	96.1	94.4	80.3	84.6	71.8	83.5	95.0	77.7
6	$L_{main}+ g$	94.7	97.5	93.3	80.8	84.6	73.1	83.1	95.0	77.1
9	**Ours** ($L_{main}+g+L_{L-Reg}$)	**94.8**	**97.6**	**93.4**	**80.8**	**84.6**	**73.2**	**83.4**	94.0	**78.0**

Table 19: MDG+GCD results: accuracy scores of each dataset. Improvements are highlighted in red.

Method	Domain gap	PACS			HomeOffice			VLCS			TerraIncognita			DomainNet		
		All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	Not	57.26	77.77	22.33	44.80	74.67	8.50	61.51	82.89	34.88	37.34	20.46	45.15	22.56	40.89	6.85
+L-Reg	minimized	55.86	77.69	19.06	43.56	71.78	9.68	61.49	81.33	36.65	40.73	29.27	35.56	25.86	47.07	7.19
Improvements		-1.40	-0.08	-3.27	-1.24	-2.89	1.18	-0.02	-1.55	1.77	3.38	8.81	-9.58	3.31	6.18	0.34
PIM	Not	56.35	71.06	27.43	43.42	72.44	8.13	63.19	80.34	40.24	47.75	35.31	50.85	24.03	42.59	7.86
+L-Reg	minimized	58.47	76.49	26.22	44.20	71.75	10.85	59.29	77.96	36.81	49.74	34.08	50.01	24.66	43.86	7.78
Improvements		2.12	5.43	-1.21	0.78	-0.70	2.72	-3.90	-2.38	-3.43	2.00	-1.23	-0.84	0.63	1.27	-0.08
MIRO	Minimized	56.83	85.62	24.85	48.28	80.61	9.03	61.53	82.72	35.03	50.22	39.92	49.45	31.49	55.44	10.57
+L-Reg		68.44	97.77	25.64	53.59	79.50	22.21	62.07	83.18	35.21	44.85	40.87	38.42	31.58	54.97	10.98
Improvements		11.61	12.14	0.79	5.31	-1.11	13.18	0.54	0.46	0.18	-5.37	0.95	-11.03	0.10	-0.47	0.41
GMDG	Sufficiently	58.33	91.46	10.18	48.85	81.41	9.22	61.36	83.31	33.75	40.02	32.38	40.07	31.15	55.17	10.18
+L-Reg	minimized	67.82	91.86	31.33	51.96	79.74	18.15	62.32	82.77	36.09	45.86	39.77	41.55	31.75	55.18	11.30
Improvements		9.50	0.40	21.15	3.11	-1.68	8.92	0.97	-0.54	2.34	5.83	7.39	1.49	0.60	0.01	1.13

Table 20: MDG+GCD results: accuracy scores of each domain in PACS dataset.

PACS	Avg			art_painting			cartoon			photo			sketch		
Method	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	57.26	77.77	22.33	47.77	90.00	0.00	56.08	83.49	20.47	59.13	47.35	68.85	66.06	90.23	0.00
with our reg	55.86	77.69	19.06	45.33	85.40	0.00	50.91	90.09	0.00	63.70	48.51	76.23	63.52	86.75	0.00
Improvements	-1.40	-0.08	-3.27	-2.44	-4.60	0.00	-5.17	6.60	-20.47	4.57	1.16	7.38	-2.54	-3.48	0.00
PIM	56.35	71.06	27.43	46.80	55.17	37.32	50.37	89.15	0.00	62.05	49.50	72.40	66.19	90.40	0.00
with our reg	58.47	76.49	26.22	46.74	88.05	0.00	56.50	78.77	27.57	64.30	48.51	77.32	66.35	90.62	0.00
Improvements	2.12	5.43	-1.21	-0.06	32.87	-37.32	6.13	-10.38	27.57	2.25	-0.99	4.92	0.16	0.22	0.00
MIRO	56.83	85.62	24.85	51.86	97.70	0.00	56.45	99.91	0.00	48.35	75.17	26.23	70.64	69.72	73.16
with our reg	68.44	97.77	25.64	68.46	97.82	35.24	61.51	98.02	14.09	72.60	98.84	50.96	71.18	96.39	2.26
Improvements	11.61	12.14	0.79	16.60	0.11	35.24	5.06	-1.89	14.09	24.25	23.68	24.73	0.54	26.67	-70.90
GMDG	58.33	91.46	10.18	51.92	97.82	0.00	54.80	96.98	0.00	56.14	74.83	40.71	70.45	96.22	0.00
with our reg	67.82	91.86	31.33	79.26	98.05	58.00	68.18	99.25	27.82	52.40	74.50	34.15	71.47	95.66	5.34
Improvements	9.50	0.40	21.15	27.33	0.23	58.00	13.38	2.26	27.82	-3.74	-0.33	-6.56	1.02	-0.56	5.34

Table 21: MDG+GCD results: accuracy scores of each domain in HomeOffice dataset.

HomeOffice	Avg			Art			Clipart			Product			Real World		
Method	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	44.80	74.67	8.50	45.26	72.68	3.26	37.94	64.48	10.19	46.71	78.74	9.87	49.28	82.80	10.68
with our reg	43.56	71.78	9.68	41.30	62.72	8.47	35.91	60.78	9.90	48.34	78.95	13.14	48.68	84.67	7.22
Improvements	-1.24	-2.89	1.18	-3.96	-9.96	5.22	-2.03	-3.70	-0.29	1.63	0.21	3.27	-0.60	1.88	-3.46
PIM	43.42	72.44	8.13	42.53	68.09	3.39	35.77	56.75	13.83	47.27	77.58	12.41	48.11	87.35	2.90
with our reg	44.20	71.75	10.85	44.64	68.85	7.56	35.48	60.90	7.49	47.49	76.32	14.35	49.17	80.92	12.59
Improvements	0.78	-0.70	2.72	2.11	0.77	4.17	-0.29	4.15	-4.92	0.23	-1.26	1.94	1.06	-6.43	9.69
MIRO	48.28	80.61	9.03	50.57	79.57	6.13	39.55	67.23	10.60	51.35	86.16	11.32	51.66	89.50	8.09
with our reg	53.59	79.50	22.21	54.02	77.87	17.47	43.87	70.98	15.52	59.94	83.95	32.12	56.54	85.21	23.52
Improvements	5.31	-1.11	13.18	3.45	-1.70	11.34	4.32	3.75	4.92	8.59	-2.21	21.00	4.88	-4.29	15.43
GMDG	48.85	81.41	9.22	51.60	81.96	5.08	40.89	69.07	11.19	51.15	87.53	9.32	51.75	86.87	11.30
with our reg	51.96	79.74	18.15	52.83	79.15	12.52	43.59	69.02	16.99	56.31	83.11	25.48	55.11	87.67	17.59
Improvements	3.11	-1.68	8.92	1.24	-2.81	7.43	2.69	-0.28	5.80	5.15	-4.42	16.16	3.36	0.80	6.30

Table 22: MDG+GCD results: accuracy scores of each domain in VLCS dataset.

VLCS	Avg			Caltech101			LabelMe			SUN09			VOC2007		
Method	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	61.51	82.89	34.88	82.07	74.87	85.85	50.54	92.01	4.85	62.07	95.15	11.38	51.35	69.51	37.45
with our reg	61.49	81.33	36.65	76.59	75.13	77.36	50.64	91.02	6.13	60.70	92.83	11.48	58.02	66.35	51.63
Improvements	-0.02	-1.55	1.77	-5.48	0.26	-8.49	0.09	-0.99	1.29	-1.37	-2.33	0.10	6.66	-3.16	14.18
PIM	63.19	80.34	40.24	80.39	72.05	84.77	53.84	91.74	12.07	62.22	94.21	13.21	56.31	63.36	50.92
with our reg	59.29	77.96	36.81	72.61	73.33	72.24	53.98	90.75	13.45	56.85	83.64	15.81	53.72	64.13	45.75
Improvements	-3.90	-2.38	-3.43	-7.77	1.28	-12.53	0.14	-0.99	1.38	-5.37	-10.57	2.60	-2.59	0.77	-5.16
MIRO	61.53	82.72	35.03	82.77	74.10	87.33	51.81	91.83	7.72	62.22	95.59	11.09	49.32	69.34	33.99
with our reg	62.07	83.18	35.21	82.51	74.62	86.66	49.51	94.43	0.00	60.97	94.65	9.35	55.31	69.00	44.84
Improvements	0.54	0.46	0.18	-0.27	0.51	-0.67	-2.31	2.60	-7.72	-1.26	-0.94	-1.74	6.00	-0.34	10.85
GMDG	61.36	83.31	33.75	82.51	74.87	86.52	49.93	95.24	0.00	59.86	93.96	7.62	53.13	69.17	40.85
with our reg	62.32	82.77	36.09	84.54	74.62	89.76	49.98	92.01	3.66	61.39	95.03	9.84	53.39	69.43	41.11
Improvements	0.97	-0.54	2.34	2.03	-0.26	3.23	0.05	-3.23	3.66	1.52	1.07	2.22	0.26	0.26	0.26

Table 23: MDG+GCD results: accuracy scores of each domain in TerraIncognita dataset.

TerraIncognita	Avg			art_painting			cartoon			photo			sketch		
Method	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	37.34	20.46	45.15	46.51	1.25	57.07	39.88	28.22	44.91	29.41	24.65	40.25	33.59	27.70	38.36
with our reg	40.73	29.27	35.56	52.94	0.14	65.27	39.26	33.76	41.64	40.24	57.45	1.03	30.47	25.71	34.32
Improvements	3.38	8.81	-9.58	6.43	-1.11	8.20	-0.62	5.53	-3.27	10.83	32.80	-39.22	-3.12	-1.99	-4.04
PIM	47.75	35.31	50.85	50.20	28.97	55.15	56.22	19.71	71.99	46.69	47.94	43.86	37.88	44.64	32.40
with our reg	49.74	34.08	50.01	53.94	34.12	58.57	59.87	19.58	77.26	47.07	62.75	11.35	38.09	19.88	52.87
Improvements	2.00	-1.23	-0.84	3.74	5.15	3.41	3.65	-0.13	5.28	0.38	14.82	-32.51	0.21	-24.76	20.47
MIRO	50.22	39.92	49.45	52.23	51.25	52.46	55.54	14.73	73.16	48.93	62.89	17.13	44.19	30.79	55.06
with our reg	44.85	40.87	38.42	56.26	22.42	64.16	31.27	23.75	34.52	54.03	65.11	28.79	37.84	52.18	26.20
Improvements	-5.37	0.95	-11.03	4.03	-28.83	11.71	-24.26	9.03	-38.64	5.10	2.22	11.66	-6.35	21.39	-28.86
GMDG	40.02	32.38	40.07	36.70	35.65	36.94	36.69	20.86	43.53	49.46	61.76	21.47	37.24	11.24	58.33
with our reg	45.86	39.77	41.55	51.89	38.16	55.09	41.30	22.05	49.61	50.47	65.29	16.72	39.77	33.59	44.79
Improvements	5.83	7.39	1.49	15.19	2.51	18.15	4.61	1.19	6.08	1.01	3.53	-4.75	2.53	22.34	-13.54

25

Table 24: MDG+GCD results: accuracy scores of each domain in DomainNet dataset.

DomainNet	Avg			clipart			info			painting		
Method	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	22.56	40.89	6.85	31.04	58.32	7.15	17.94	34.71	6.85	30.59	51.82	9.34
with our reg	25.86	47.07	7.19	32.03	58.43	8.91	18.17	34.31	7.50	31.93	52.58	11.24
Improvements	3.31	6.18	0.34	0.99	0.11	1.76	0.23	-0.41	0.65	1.33	0.76	1.90
PIM	24.03	42.59	7.86	32.01	57.38	9.80	18.80	33.56	9.03	22.22	36.62	7.80
with our reg	24.66	43.86	7.78	31.91	57.76	9.26	16.99	30.77	7.89	28.17	45.94	10.37
Improvements	0.63	1.27	-0.08	-0.11	0.38	-0.54	-1.80	-2.79	-1.15	5.95	9.32	2.57
MIRO	31.49	55.44	10.57	40.13	67.55	16.11	25.84	48.53	10.84	37.89	62.45	13.29
with our reg	31.58	54.97	10.98	40.61	66.72	17.75	25.58	45.83	12.19	36.74	62.29	11.15
Improvements	0.10	-0.47	0.41	0.49	-0.83	1.64	-0.26	-2.70	1.35	-1.15	-0.16	-2.14
GMDG	31.15	55.17	10.18	40.38	70.69	13.84	24.96	46.50	10.72	36.29	59.80	12.75
with our reg	31.75	55.18	11.30	40.91	68.17	17.05	26.60	49.11	11.71	36.82	60.76	12.85
Improvements	0.60	0.01	1.13	0.53	-2.52	3.21	1.63	2.61	0.99	0.53	0.96	0.10

| DomainNet	Avg			quickdraw			real			sketch		
Method	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown	All	Known	Unknown
ERM	-	-	-	8.88	12.83	4.91	17.88	31.20	4.10	29.01	56.45	8.76
with our reg	-	-	-	9.04	14.73	3.31	34.34	63.94	3.69	29.68	58.41	8.49
Improvements	-	-	-	0.16	1.91	-1.59	16.45	32.74	-0.41	0.67	1.96	-0.27
PIM	-	-	-	9.92	14.73	5.09	29.09	53.88	3.42	32.12	59.35	12.03
with our reg	-	-	-	9.94	15.11	4.74	30.26	56.13	3.47	30.68	57.43	10.95
Improvements	-	-	-	0.02	0.38	-0.35	1.17	2.25	0.05	-1.44	-1.93	-1.08
MIRO	-	-	-	8.06	12.12	3.98	42.19	75.49	7.72	34.83	66.51	11.46
with our reg	-	-	-	9.36	15.73	2.95	42.00	74.36	8.50	35.23	64.89	13.34
Improvements	-	-	-	1.30	3.61	-1.03	-0.20	-1.14	0.78	0.40	-1.62	1.88
GMDG	-	-	-	7.43	11.83	3.01	42.84	75.27	9.27	35.01	66.95	11.46
with our reg	-	-	-	9.11	13.51	4.70	42.63	74.42	9.72	34.44	65.13	11.80
Improvements	-	-	-	1.68	1.67	1.69	-0.21	-0.84	0.45	-0.58	-1.81	0.34
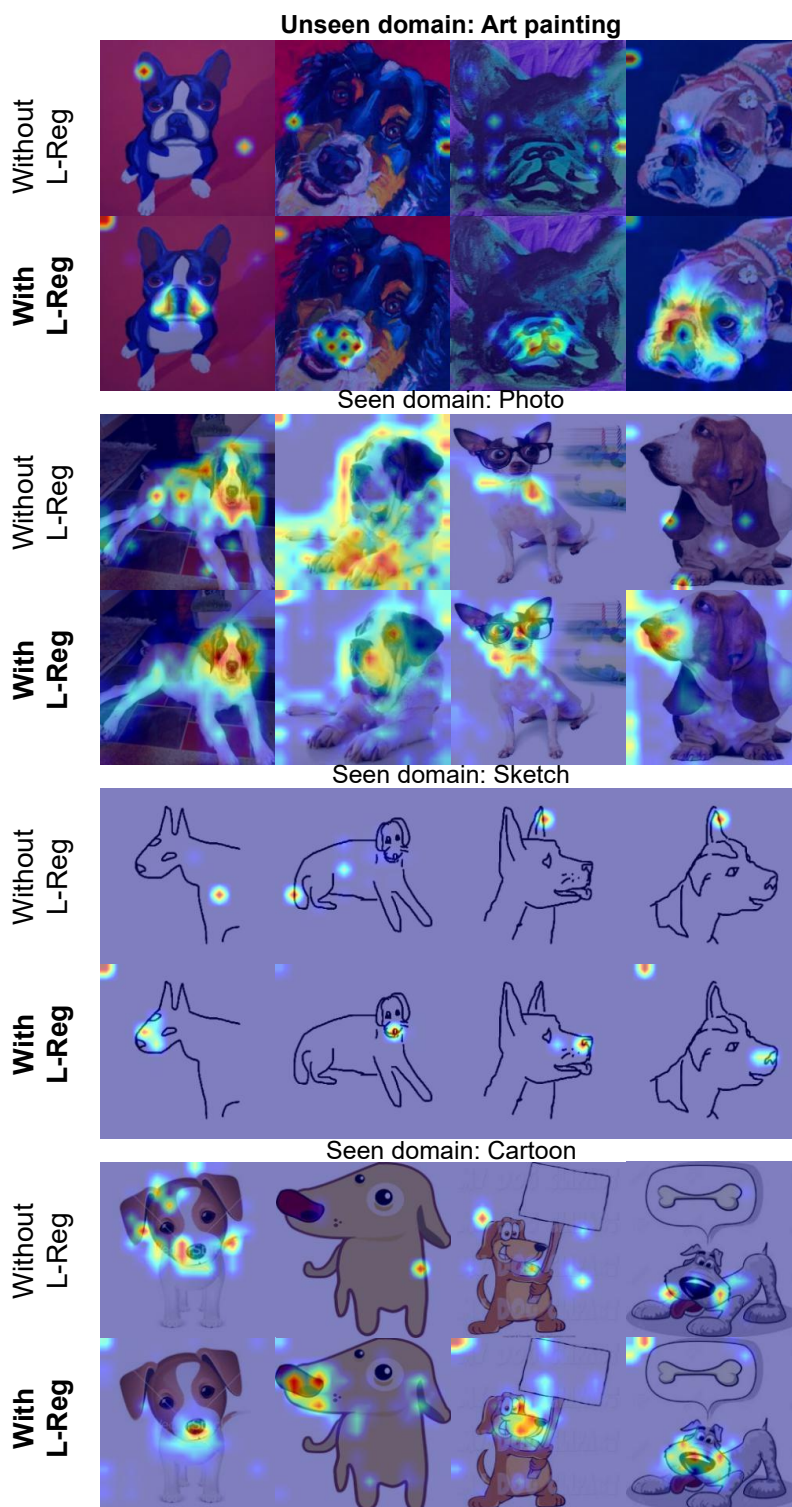
Figure 7: GradCAM visualizations: Baseline is GMDG. The used dataset is PACS. The model is trained under uDG+GCD setting with and without L-Reg, respectively. It can be seen that for the **known** class 'dog,' the model trained with L-Reg extracts the area around the nose area for classification across all seen and unseen domains.
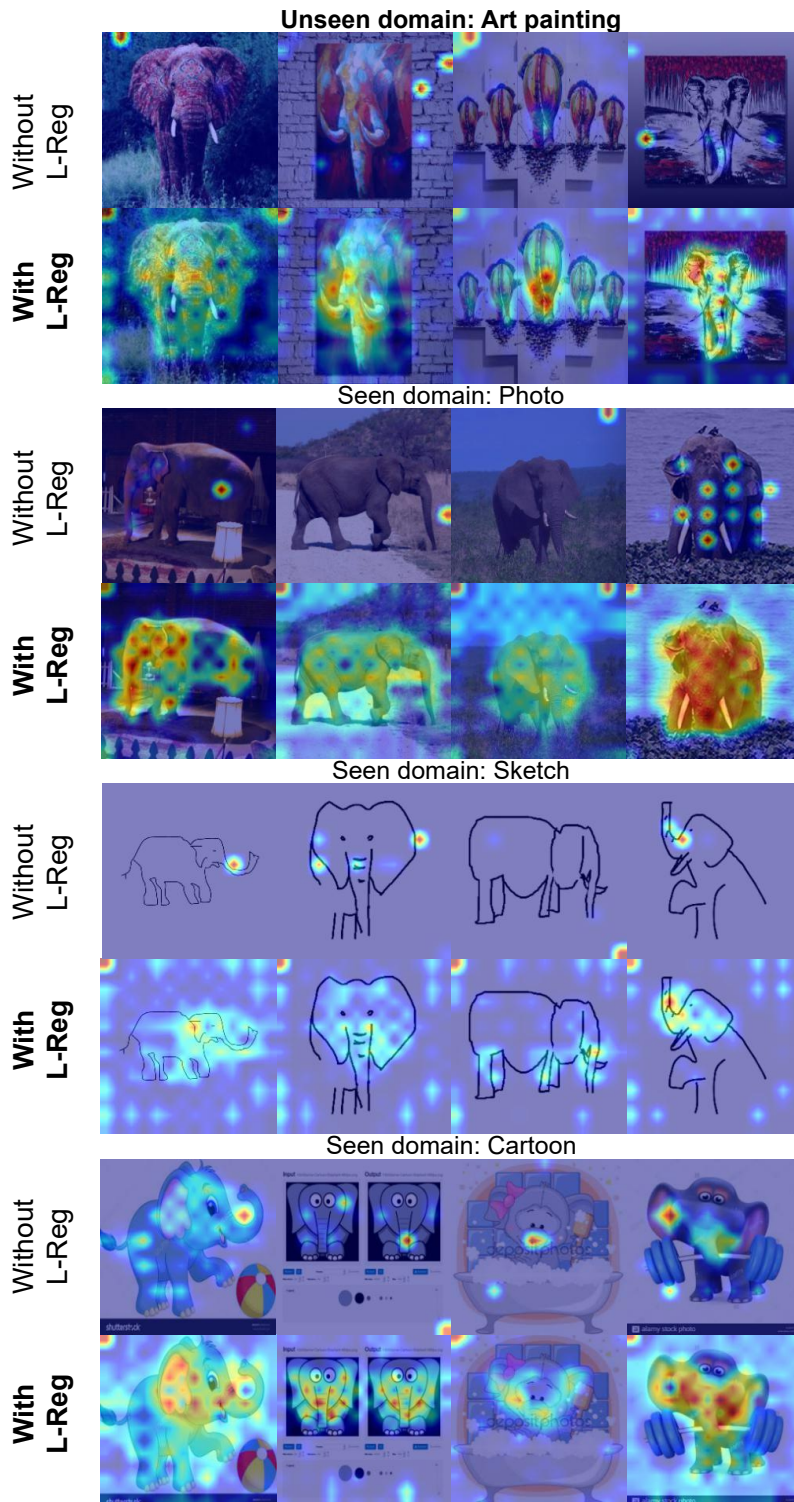
Figure 8: GradCAM visualizations: Baseline is GMDG. The used dataset is PACS. The model is trained under uDG+GCD setting with and without L-Reg, respectively. It can be seen that for the **known** class 'elephant,' the model trained with L-Reg extracts the shape of long noses, teeth, and big ears for classification across all seen and unseen domains. The compromise of the known sets can be seen in the sketch domain, where those features are not significant.
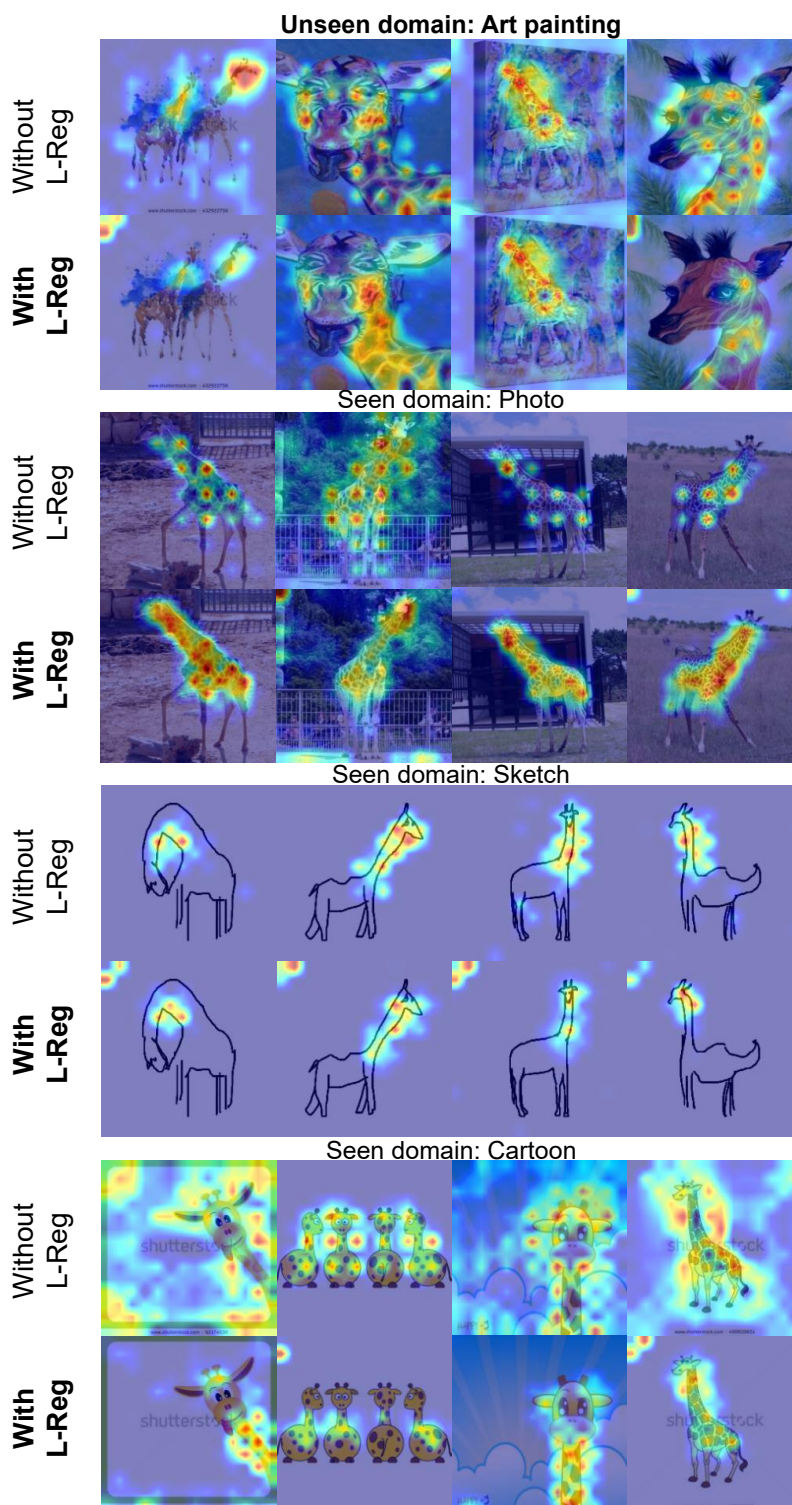
Figure 9: GradCAM visualizations: Baseline is GMDG. The used dataset is PACS. The model is trained under uDG+GCD setting with and without L-Reg, respectively. It can be seen that for the **known** class 'giraffe,' the model trained with L-Reg extracts the feature of the long necks for classifying across all seen and unseen domains.
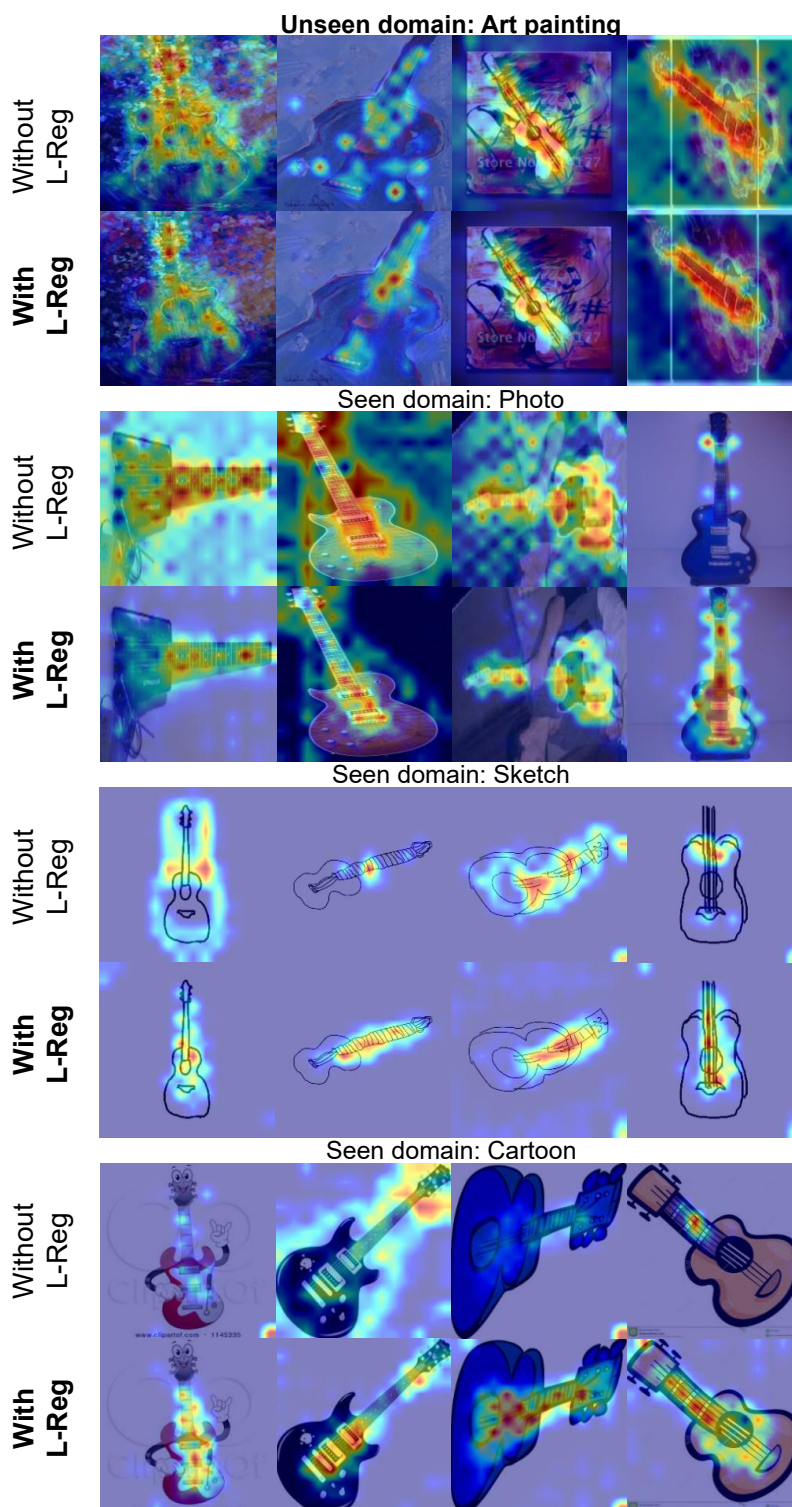
Figure 10: GradCAM visualizations: Baseline is GMDG. The used dataset is PACS. The model is trained under uDG+GCD setting with and without L-Reg, respectively. It can be seen that for the **known** class 'guitar,' the model trained with L-Reg extracts the features of the necks and the strings of the guitar for classification across all seen and unseen domains.
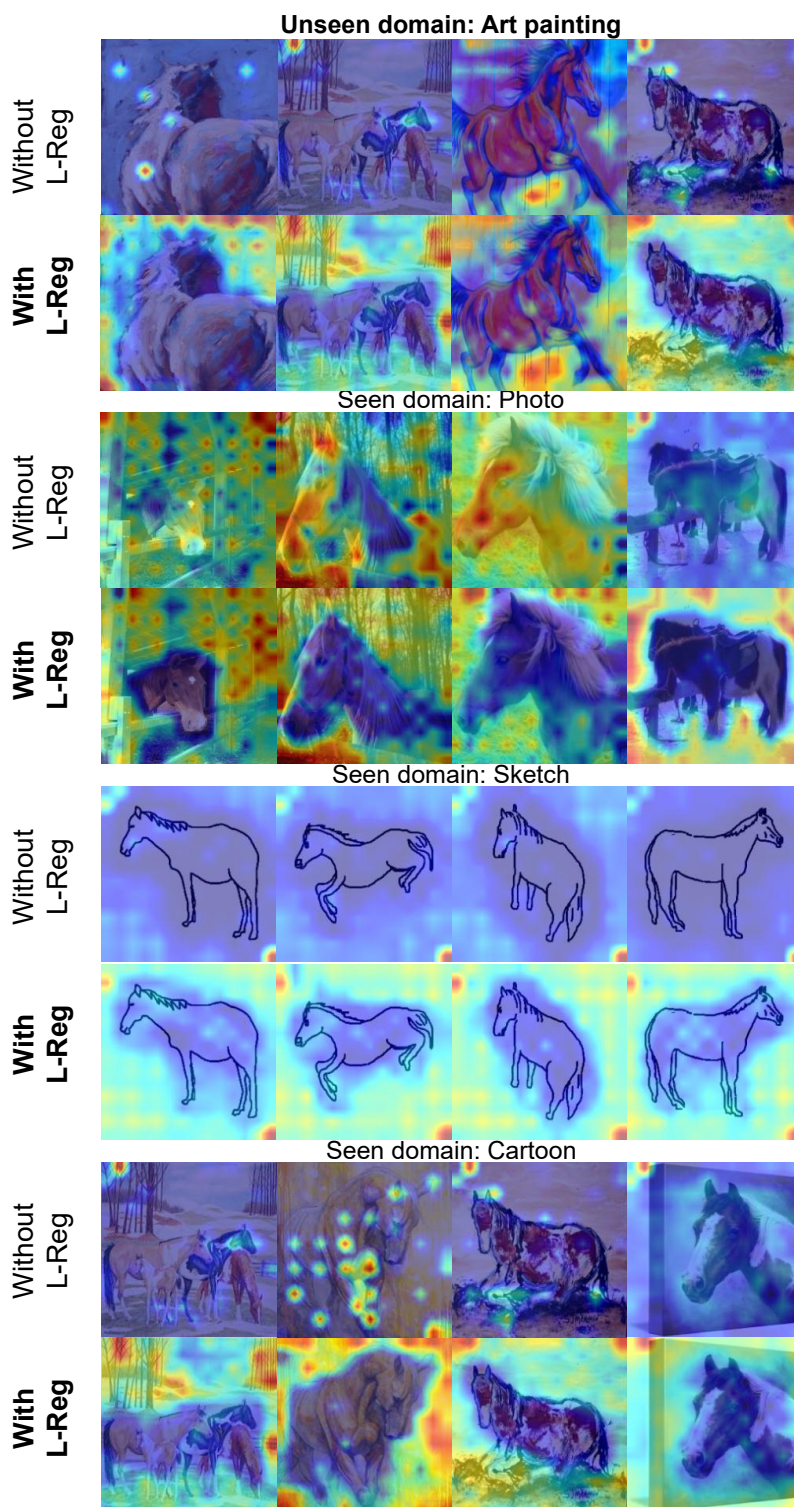
Figure 11: GradCAM visualizations: Baseline is GMDG. The used dataset is PACS. The model is trained under uDG+GCD setting with and without L-Reg, respectively. It can be seen that for the **unknown** class 'horse,' the model trained with L-Reg extracts the features of the overall outline shapes of horses for classification across all seen and unseen domains.
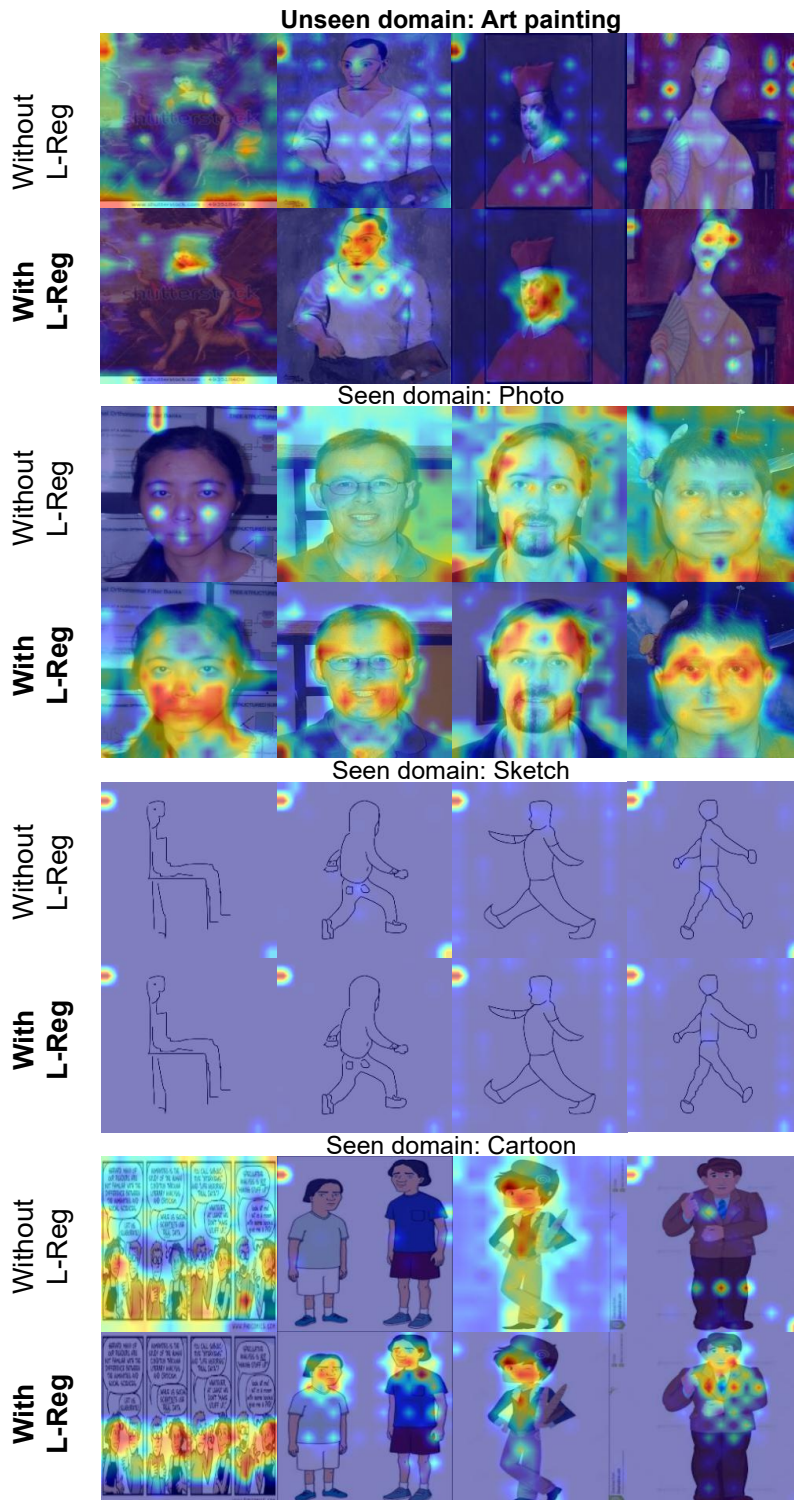
Figure 12: GradCAM visualizations: Baseline is GMDG. The used dataset is PACS. The model is trained under uDG+GCD setting with and without L-Reg, respectively. It can be seen that for the **unknown** class 'person,' the model trained with L-Reg extracts the features of human faces for classification across all seen and unseen domains. The compromise of the known sets can be seen in the sketch domain, where those faces are not drawn.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the paper discusses the limitations of the work in the Conclusion Section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Yes, the paper provides the full set of assumptions and a complete proof in the main paper and appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: This paper provides all the information needed to reproduce the main experimental results of the paper as much as possible in the appendix and supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides uses open-source data and open access to the code. The code is especially attached as supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, this paper specifies all the training and test details in the Experiments Section and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Following previous work, this paper does not report error bars. However, this paper reports some results in the form of mean with standard deviation, as shown in Table 1. However,

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides information on the computer resources in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Yes, the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We think that there is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper seems to pose no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets (e.g., code, data, models) used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: New assets introduced in the paper are documented, and the documentation is provided alongside the assets. The model training code is provided in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.