Bridging the Domain Gap by Clustering-based Image-Text Graph Matching

Nokyung Park [1] **Daewon Chae** [1] **Jeong Yong Shim** [2] **Sangpil Kim** [3] **Eun-Sol Kim** [*4] **Jinkyu Kim** [*1]

Abstract

Learning domain-invariant representations is important to train a model that can generalize well to unseen domains. To this end, we propose a novel approach that leverages the semantic structures inherent in text descriptions as effective pivot embeddings for domain generalization. Specifically, we utilize graph representations of images and their associated textual descriptions to obtain domain-invariant pivot embeddings that capture the underlying semantic relationships between local images and text descriptors. Our approach involves a clustering-based graph-matching algorithm that matches graph-based image node features into textual graphs. Experimental results show the efficacy of our proposed method in enhancing the generalization ability of the model.

1. Introduction

Domain generalization aims at improving a model's generalization ability for unseen domains. Existing domain generalization approaches involve a range of techniques, including reducing domain discrepancies (Sun & Saenko, 2016) and implementing data augmentation (Yan et al., 2020). Other studies have explored using auxiliary semantic cues to learn domain-invariant features (Cha et al., 2022). A recent approach (Min et al., 2022) utilizes text descriptions as auxiliary cues to obtain domain-invariant features.

In this paper, we suggest using multimodal graph representations to get effective domain-invariant pivot embeddings for domain generalization problems. Our method represents text descriptions and images in their respective graphs and aligns their embeddings by matching the graphs. By grounding image features into textual graphs, which capture human reasoning, we can learn robust domain-invariant features. Our approach facilitates multilevel semantic alignment by clustering node features and matching multimodal graphs. Our method achieves state-of-the-art or superior performance on two benchmark datasets, CUB-DG (Min et al., 2022) and DomainBed (Gulrajani & Lopez-Paz, 2020).

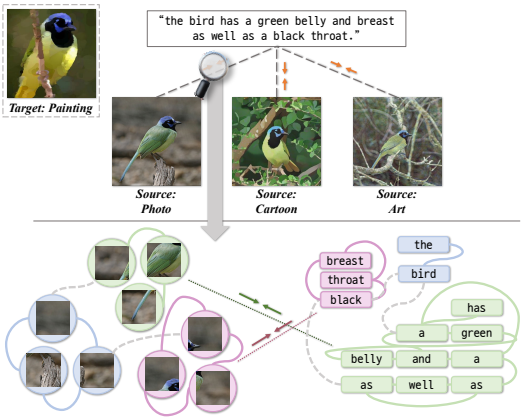


Figure 1. Our model uses clustering-based graph modeling for visual and textual features to match clusters, yielding domain-invariant features for better generalization in unseen domains.

Our contributions can be summarized as follows.

- We propose a novel approach that utilizes graph representations for both image and text inputs for domain generalization problem.
- We introduce a method that aligns multimodal graphs using a graph neural network, clustering and matching node features.
- Our approach achieves state-of-the-art performance on the CUB-DG benchmark dataset.

2. Method

Given a distribution over multiple (or single) source domains $\{\mathcal{S}_1, \mathcal{S}_2, \dots\} \in \mathcal{S}$, the domain generalization problem considers the following classical stochastic optimization, in which we minimize the data-dependent generalization upper bound of the expected task loss (Sinha et al., 2017):

$$\underset{\theta}{\text{minimize}} \sup_{\mathcal{T}:\mathcal{D}(\mathcal{S},\mathcal{T})\leq\rho} \mathbb{E}_{\mathcal{T}}\big[\mathcal{L}(\theta;\mathcal{S})\big] \qquad (1)$$

[1]Department of Computer Science and Engineering, Korea University [2]Department of Artificial Intelligence Application, Hanyang University [3]Department of Artificial Intelligence, Korea University [4]Department of Computer Science, Hanyang University. [*]Correspondence to: Jinkyu Kim <jinkyukim@korea.ac.kr>, Eun-Sol Kim <eunsolkim@hanyang.ac.kr>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2023 by the author(s).

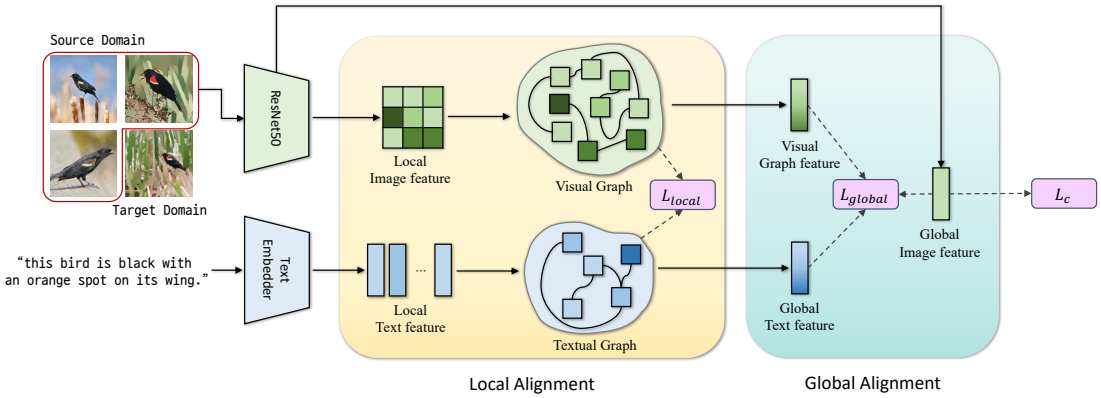


Figure 2. An overview of our proposed method.

where we consider unseen target domains $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots\}$ and the discrepancy between $\mathcal{S}$ and $\mathcal{T}$ is bounded by an arbitrary bound ρ. In this paper, we consider image classification scenarios.

Inspired by recent work (Min et al., 2022), but unlike this, our proposed model learns to extract visual representations that align not only globally but also locally with explicitly verbalized knowledge from human reasoning.

Our model consists of three main parts: (i) a Graph-based Visual Encoder, (ii) a Graph-based Textual Encoder, and (iii) Local and Global Alignment between Visual and Textual Graphs. Our approach is shown in Figure 2.

Following standards in the domain generalization task, we use the backbone ResNet50 (He et al., 2016b) to take images $\mathcal{I}$ as an input, outputting a d-dimensional global visual representation $\mathbf{x}_g \in \mathbb{R}^d$. This global representation $\mathbf{x}_g$ is trained to predict its classification label $\hat{\mathbf{y}}$ with a linear layer. Our backbone and a classifier are trained by a standard cross-entropy loss L_c.

2.1. Graph-based Visual Encoder

We construct a graph with visual representations to enhance the model's generalizability to unseen domain. Given M number of d-dimensional *local* visual representations $\mathbf{x}_l \in \{x_{l,1}, x_{l,2}, \dots, x_{l,M}\}$ extracted from intermediate layers of the backbone, we consider these representations as a set of unordered nodes. Inspired by recent work (Han et al., 2022), we construct a visual graph $\mathcal{G}_v$ with M nodes, and connect each node to its K_v nearest neighbors by using L_2 distance.

Given the visual graph $\mathcal{G}_v$, we further apply two layers of graph convolution network (GCN) (Kipf & Welling, 2016) followed by a linear and a BatchNorm (Ioffe & Szegedy, 2015) layers to learn relational knowledge between local visual representations. This results in $\mathbf{g}_v \in \mathbb{R}^{d_g}$, the final locally-aware visual graph representation. Note that, we also add an additional classifier that takes the $\mathbf{g}_v$ as an input to create a graph that better captures the characteristics of the class. We provide detailed explanations in the Appendix.

2.2. Graph-based Textual Encoder

We build a textual graph from a natural language description of each class, followed by aligning both visual and textual graphs to learn domain-invariant visual representations.

A sequence of L (at maximum) words is tokenized and encoded with a standard word-level (learnable) embedding layer, producing d_t-dimensional embedding vectors $\mathbf{t} \in \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_L\}$ where $\mathbf{t}_i \in \mathbb{R}^{d_t}$. The Textual Graph $\mathcal{G}_t$ is constructed similarly to the Visual Graph, connecting each node $\mathbf{t}_i$ to its K_t nearest neighbors based on the smallest L_2 distance. We then apply a GCN network, which has the same architecture described in Section D, on $\mathcal{G}_t$ to obtain a textual graph representation $\mathbf{g}_t \in \mathbb{R}^{d_g}$.

2.3. Local and Global Alignment between Visual and Textual Graphs

We apply the following two graph-matching approaches: (i) Global Graph Matching and (ii) Clustering-based Fine-grained Graph Matching.

Global Graph Matching. A standard approach to matching two different graph representations is minimizing the Euclidean distance as follows:

$$\begin{aligned}\mathcal{L}_{\text{global}} = &||f_{\text{proj},x}(\mathbf{x}_g) - f_{\text{proj},v}(\mathbf{g}_v)||_2 \\ &+||f_{\text{proj},x}(\mathbf{x}_g) - f_{\text{proj},t}(\mathbf{g}_t)||_2\end{aligned} \tag{2}$$

where we use a linear layer to project each feature (i.e. $\mathbf{x}_g$, $\mathbf{g}_v$, and $\mathbf{g}_t$) such that these three projected features are pulled together. Note that $f_{\text{proj},x}$, $f_{\text{proj},v}$, and $f_{\text{proj},t}$ represent the respective projection layers. To avoid representation collapse while pulling latent representations together, we add an auxiliary classifier that takes $f_{\text{proj},x}(\mathbf{x}_g)$ as input and outputs per-class probabilities. It is trained with standard cross-entropy loss.

Clustering Graph Nodes. We also use local graph matching to align node-level features of the image and text graphs based on similar semantic cues. To ensure the two graphs have the same level of semantic representation despite different node representations, we apply a graph clustering al-

gorithm. Our method defines user-specified parameters N_v and N_t for the number of clusters in the visual and textual graphs respectively. Note that we set $N_v \geq N_t$ since images may contain visual contents (e.g. backgrounds) that are not generally described in the text. We use a modularity-based graph clustering method (Tsitsulin et al., 2020) to reflect the semantic structure of the graph, which is constructed based on node similarity.

Graph Cluster Matching. We use the set-based loss, i.e. bipartite matching loss, between two disjoint sets of clusters: (i) a set of clusters $\mathcal{C}_v \in \{\mathcal{C}_v^1, \mathcal{C}_v^2, \ldots, \mathcal{C}_v^{N_v}\}$ of the visual graph $\mathcal{G}_v'$ and (ii) a set of clusters $\mathcal{C}_t \in \{\mathcal{C}_t^1, \mathcal{C}_t^2, \ldots, \mathcal{C}_t^{N_t}\}$ from the textual graph $\mathcal{G}_t'$. We minimize the following pair-wise matching loss:

$$\mathcal{L}_p = \frac{1}{N_t} \sum_{i=1}^{N_t} ||\mathcal{C}_v^{\mu_i} - \mathcal{C}_t^i||_2 \tag{3}$$

where $\mu_i \in \{1, 2, \ldots, N_v\}$ is the node index of the cluster in $\mathcal{C}_v$ which matches to i in $\mathcal{C}_t$, producing the smallest total Euclidean distance by bipartite matching.

To prevent representation collapse, we use a hinge loss based on negative pairs formed by cluster representations $\mathcal{C}_v^i$ and $\mathcal{C}'^j_t$ from different input images. The matched closest distance $||\mathcal{C}_v^{\mu_i} - \mathcal{C}_t^i||_2$ should be smaller than any other pairs between $\mathcal{C}_v^j$ and $\mathcal{C}'^i_t$ (or $\mathcal{C}'^j_v$ and $\mathcal{C}_t^i$).

$$\begin{aligned}\mathcal{L}_h = &\max(0, \mathcal{L}_p - \text{MinDist}(\mathcal{C}'_v, \mathcal{C}_t) + \epsilon) \\ &+\max(0, \mathcal{L}_p - \text{MinDist}(\mathcal{C}_v, \mathcal{C}'_t) + \epsilon)\end{aligned} \tag{4}$$

where $\text{MinDist}(\mathcal{C}'_v, \mathcal{C}_t)$ represents the minimum pair-wise matching loss similar to $\mathcal{L}_p$, but is applied to different inputs within a mini-batch. We compute it over all pairs of samples in a mini-batch and use the average as the final loss value:

$$\mathcal{L}_{\text{local}} = \frac{1}{B} \sum_b (\lambda_p \mathcal{L}_p + \lambda_h \mathcal{L}_h + \lambda_d \mathcal{L}_d) \tag{5}$$

where L_d is clustering loss which is defined in (Tsitsulin et al., 2020). Note that λ_p, λ_h, and λ_d are hyper-parameters that control the weight of each loss term. We set the size of a mini-batch to B. We also add an auxiliary classifier that takes the average-pooled cluster representation $\frac{1}{N_t}\sum_{i=1}^{N_t} \mathcal{C}_v^{\mu_i}$ as an input and outputs the per-class softmax probability, trained with the standard cross-entropy loss.

Loss Function. Ultimately, we train our model end-to-end by minimizing the following loss L:

$$\mathcal{L} = \mathcal{L}_c + \lambda_{\text{global}} \mathcal{L}_{\text{global}} + \lambda_{\text{local}} \mathcal{L}_{\text{local}} \tag{6}$$

where λ_{global}, and λ_{local} are hyper-parameters to control the strength of each loss term.

3. Experiments

3.1. Datasets

We use CUB-DG, an extension of the CUB dataset (Welinder et al., 2010) with up to 10 text descriptions per image, for domain generalization task. The dataset is split into four domains: Photo, Cartoon, Art, and Painting. We follow the common experimental protocol and use the official split. We also evaluate our model on four datasets from DomainBed (Gulrajani & Lopez-Paz, 2020). Especially on the following four datasets: VLCS (Fang et al., 2013), PACS (Venkateswara et al., 2017), OfficeHome (Li et al., 2017) and TerraIncognita (Beery et al., 2018). We use the class definitions from Oxford dictionary for text inputs as in GVRT.

Table 1. Results (in %) on CUB-DG dataset with multi-source DG setting. *Abbr.* I: Image, T: Text.

Algorithm	Modality	Target Domain				Avg.
		Photo	Cartoon	Art	Paint	
Ours	I+T	**75.4**	**65.5**	**54.0**	**41.4**	**59.1**
GVRT (PTE) (Min et al., 2022)	I+T	74.6	64.2	52.2	37.0	57.0
CORAL (Sun & Saenko, 2016)	I	72.2	63.5	50.3	35.8	55.4
SD (Pezeshki et al., 2020)	I	71.3	62.2	50.8	34.8	54.7
SagNet (Nam et al., 2021)	I	67.4	60.7	44.0	34.2	51.6
Mixup (Yan et al., 2020)	I	67.1	55.9	51.1	27.2	50.3
DANN (Ganin et al., 2016)	I	67.5	57.0	42.8	30.6	49.5
VREx (Krueger et al., 2020)	I	63.9	54.9	38.6	30.1	46.9
ERM (Vapnik, 1999)	I	62.5	53.2	37.4	29.0	45.5

Table 2. Results (in %) on the Domainbed with multi-source DG setting.

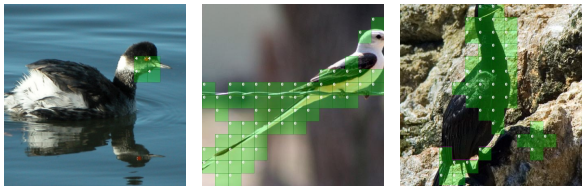
Algorithm	Dataset				Avg.
	VLCS	PACS	OfficeHome	TerraIncognita	
Ours	78.3 ± 0.4	85.7 ± 0.1	70.1 ± 0.1	49.5 ± 0.9	70.9
GVRT (PTE) (Min et al., 2022)	79.0 ± 0.2	85.1 ± 0.3	70.1 ± 0.1	48.0 ± 0.2	70.6
MIRO (Cha et al., 2022)	79.0 ± 0.0	85.4 ± 0.4	70.5 ± 0.4	50.4 ± 1.1	71.3
CORAL (Sun & Saenko, 2016)	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	70.3
SagNet (Nam et al., 2021)	77.8 ± 0.5	86.3 ± 0.2	68.1 ± 0.1	48.6 ± 1.0	70.2
SelfReg (Kim et al., 2021)	77.8 ± 0.9	85.6 ± 0.4	67.9 ± 0.7	47.0 ± 0.3	69.6
Mixup (Yan et al., 2020)	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	69.5
ERM (Vapnik, 1999)	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	68.9

3.2. Results

We evaluate our model by using a single domain as the target and the others as sources. Our model is evaluated on CUB-DG and compared with 13 other DG algorithms, as shown in Table 1. The complete table is in the Appendix G. Our proposed method clearly outperforms the other domain generalization techniques in all target domains with a significant gain. In terms of the average image classification accuracy, ours show 59.1%, which is 2.1% higher than GVRT (which uses the same image and text inputs) and 3.7%-14.3% higher than other image-only approaches.

In Table 2, our method ranks 2nd in average performance among the top 11 algorithms among the top 11 algorithms. The complete table is also in the Appendix G. However, we believe that there is still room for improvement, as textual definitions often include non-visual descriptions that limit the benefits of using a multi-modal dataset.

Figure 3 shows image regions and sets of words that are matched by our model, such as matching a bird's eye re-



(a) (b) (c)

(a) This bird **has** wings that are black **and has orange eyes**.

(b) This bird **has a** white head **and chest** that **slowly turns in to a peach color near his feet, and has extremely long tail** feathers that are **longer than** his body.

(c) This bird **has** wings that are black **and has a long neck**.

Figure 3. Examples of the image region in visual graph clusters matched with texts in textual graph clusters.

gion with the words "orange eyes". Our model effectively matches image features with class-discriminative texts.

Ablation studies and more analyses are in Appendix. Furthermore, robust qualitative visualizations in Appendix suggest that our model learns domain-invariant features.

4. Conclusion

We propose a novel method, which utilizes textual descriptions by aligning them with a clustering-based graph-matching algorithm to train domain-invariant visual representations. We evaluate our model with state-of-the-art domain generalization approaches on CUB-DG and DomainBed datasets, achieving higher or matched scores than baselines.

Acknowledgements. This work was supported by the National Research Foundation of Korea grant (NRF-2021R1C1C1009608, 10%), Basic Science Research Program (NRF-2021R1A6A1A13044830, 10%) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00264 (45%) and 2022-0-00043 (25%)). S. Kim is partially supported by Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023 (Project Name: 4D Content Generation and Copyright Protection with Artificial Intelligence, Project Number: R2022020068, 10%).

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021.

Bui, M.-H., Tran, T., Tran, A., and Phung, D. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021.

Cha, J., Lee, K., Park, S., and Chun, S. Domain generalization by mutual-information regularization with pre-trained models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pp. 440–457. Springer, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Fang, C., Xu, Y., and Rockmore, D. N. 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 (ICCV)*, pp. 1657–1664, 2013.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. *CoRR*, abs/1704.01212, 2017. URL http://arxiv.org/abs/1704.01212.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Han, K., Wang, Y., Guo, J., Tang, Y., and Wu, E. Vision gnn: An image is worth graph of nodes. *arXiv preprint arXiv:2206.00272*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016b.

Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 124–140. Springer, 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Kim, D., Yoo, Y., Park, S., Kim, J., and Lee, J. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018b.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018c.

Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., and Zhang, Y. Graph structured network for image-text matching. *CoRR*, abs/2004.00277, 2020. URL https://arxiv.org/abs/2004.00277.

Min, S., Park, N., Kim, S., Park, S., and Kim, J. Grounding visual representations with texts for domain generalization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pp. 37–53. Springer, 2022.

Nam, H. et al. Reducing domain gap by reducing style bias. In *CVPR*, 2021.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *arXiv preprint arXiv:2011.09468*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *ICLR*, 2017.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 443–450. Springer, 2016.

Tsitsulin, A., Palowitch, J., Perozzi, B., and Müller, E. Graph clustering with graph neural networks. *arXiv preprint arXiv:2006.16904*, 2020.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10 (5):988–999, 1999.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Yan, S., Song, H., Li, N., Zou, L., and Ren, L. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

Zhang, M., Marklund, H., Gupta, A., Levine, S., and Finn, C. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020.

Zhou, K., Yang, Y., Qiao, Y., and Xiang, T. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020.

A. Overview

This appendix contains supplementary information that couldn't be included in the main manuscript due to page constraints. Section B provides an overview of the related works, and Section C covers implementation details. Additionally, Section D offers detailed information on the graph-based visual representation. Furthermore, this appendix includes elaborate analyses in Sections E and F. Finally, Section H presents the detailed results of the Domainbed experiments.

B. Related Works

B.1. Domain Generalization.

Domain generalization refers to the task of improving a model's generalization performance on unseen target domains where data distribution differs from the source domains. The main idea of domain generalization is to learn domain-invariant features from multiple source domains. Various methods have been proposed to resolve this problem by (i) reducing domain discrepancies in the feature space (Sun & Saenko, 2016; Ganin et al., 2016), (ii) by implementing data augmentation (Yan et al., 2020). (iii) Other studies have proposed using auxiliary semantic cues to facilitate learning domain-invariant features (Cha et al., 2022).

Recently, GVRT (Min et al., 2022) successfully leverages textual descriptions for models to learn domain-invariant visual representations by aligning them with verbalized (domain-invariant and class-discriminative) knowledge from humans' typical reasoning (e.g., given a text "this bird is black with an orange spot on its wing"). Following the similar line of GVRT (Min et al., 2022), we also want to improve the model's generalization power by leveraging visual and textual inputs together. However, we focus more on aligning locally-aware high-order semantic relations via graph structures instead of simply matching global representations.

B.2. Graph Neural Network.

Along with the huge success of neural networks in computer vision and natural language processing domains, new methodologies to deal with irregular structural inputs have been recently suggested. To learn the representations from the structural inputs, such as molecular graphs, social networks, and meshes, various types of graph-based neural network algorithms are suggested. According to the ways of representing graph data, convolution-based methods(GCN (Kipf & Welling, 2016)), attention-based methods(GAT (Veličković et al., 2017)), and message-passing methods(MPNN (Gilmer et al., 2017)) can be applied to graph representation learning. Recently, the applications of graph neural networks have been extended to image and text domains (Liu et al., 2020). By representing the image and text inputs as graphs, it becomes possible to consider the irregular and high-order correlations between tokens. In this paper, we suggest representing the multimodal inputs as graphs and matching the semantic correspondences between the multimodal inputs using graph neural networks to get the domain-invariant features.

C. Implementation Details

Same as previous domain generalization approaches, we also use ImageNet (Deng et al., 2009) pre-trained ResNet-50 (He et al., 2016a) as our backbone, yielding a 2,048-dimensional visual representation from the last layer. Our model is trained end-to-end for 5,000 training steps using Adam optimizer with a learning rate of 5e-5. We use the value of 0.1, 0.1, and 1 for λ_p, λ_h and λ_d. For training, we use standard image augmentations techniques such as random cropping, horizontal flipping, color jittering, grayscale conversion, and normalization. Our implementation is based on DomainBed (Gulrajani & Lopez-Paz, 2020), which is a unified domain generalization testbed. For hyperparameters, we set a batch size to 32 for each source domain, and we use both 1 for λ_{global} and λ_{local}, respectively.

D. Locally-aware Visual Graph Construction

In this section, we describe more detailed process of constructing the locally-aware visual graph. First, our backbone (ResNet50 (He et al., 2016b)) produces features of size $m' \times m' \times d$. The features are then transformed through average pooling to obtain a size of $m \times m \times d$, where $m \times m$ corresponds to M. This pooling operation is equivalent to dividing the image $\mathcal{I}$ into M grids (refer to Figure 4 (a)). Each grid corresponds to a node in the visual graph, and possesses a d-dimensional feature. In our experiments, we set m and M to 14 and 196, respectively. Next, we compute the L_2 distance

between each node and all other nodes in the graph, and sort them in ascending order. Subsequently, we select the K_v nearest nodes to each node. Figure 4 (b) shows the process of ranking nodes based on their L_2 distance from each node, with only the top two nodes selected. Finally, we can build the locally-aware visual graph, which has M nodes with K_v neighboring nodes.

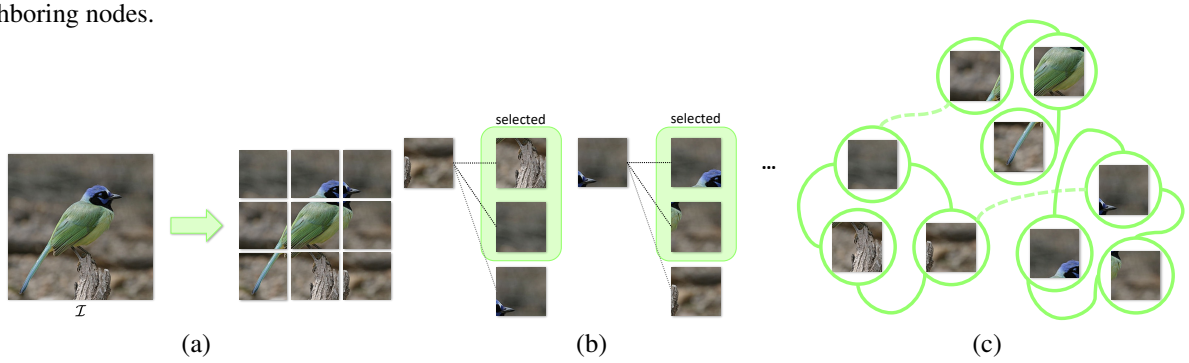


Figure 4. 3 Steps to construct Locally-aware Visual Graph

E. Ablation Study

Table 3. Out-of-distribution test accuracies (in %) on the CUB-DG dataset. We compare variants of our model with and without (i) Visual Graph and (ii) Textual Graph.

Visual Graph	Textual Graph	Target Domain				Avg.
		Photo	Cartoon	Art	Paint	
-	-	68.5	59.0	38.6	32.5	49.6
✓	-	70.3	57.0	48.1	33.5	52.2
-	✓	75.0	64.4	53.0	34.7	56.8
✓	✓	**75.4**	**65.5**	**54.0**	**41.5**	**59.1**

Table 4. Out-of-distribution test accuracies (in %) on the CUB-DG dataset. We compare variants of our model with and without (i) global graph alignment (i.e. graph-level feature matching) and (ii) local graph alignment (i.e. clustering-based graph node matching).

Global Alignment $\mathcal{L}_{\text{global}}$	Local Alignment $\mathcal{L}_{\text{local}}$	Target Domain				Avg.
		Photo	Cartoon	Art	Paint	
-	-	65.1	52.5	38.2	29.0	46.2
-	✓	71.4	61.3	49.4	34.5	57.2
✓	✓	**75.4**	**65.5**	**54.0**	**41.5**	**59.1**

E.1. Two types of Graphs

We analyze the effect of graph-based visual and textual encoders regarding the out-of-distribution test accuracies. Table 3 shows that either using visual or textual graph alone improves model generalization, but the gain is marginal with the visual graph alone. Also, the gain is maximized by using both graphs. This indicates that building a graph structure effectively transfers text knowledge to train a generalizable visual encoder.

E.2. Two types of Alignment Losses

We also perform an ablation study to verify the effect of our proposed global and local alignments. As shown in Table 4, using a global alignment, which aligns graph-level features together, is effective in improving accuracies, especially in photo, cartoon, and art domains. Adding local alignment, which aligns graphs via the clustering-based matching algorithm, improves all domains, especially a large gain in the Paint domain is observed. Overall, a model using global and local alignments generally outperforms the alternatives, showing 59.1% in average accuracy.

Table 5. Out-of-distribution test accuracies (in %) on the CUB-DG dataset. We compare our model with and without graph-based visual representation classification.

Classification for $\mathbf{g}_v$	Target Domain				Avg.
	Photo	Cartoon	Art	Paint	
-	74.7	62.3	52.3	35.7	56.2
✓	**75.4**	**65.5**	**54.0**	**41.5**	**59.1**

E.3. Graph-based Visual Representation Classification

As described in the paper, we introduced an additional classifier that takes $\mathbf{g}_v$ as an input to effectively capture the class-discriminative features. This classifier is a linear layer trained by the standard cross-entropy loss. Analysis of the results presented in Table 5 shows the inferior performance when the aforementioned classifier is not trained, demonstrating that the classifier is crucial to performance.

F. Visualization

F.1. GradCAM Visualizations

(a) GVRT (b) GDG

Figure 5. GradCAM (Selvaraju et al., 2017) visualizations to evaluate where the model sees for (a) GVRT (Min et al., 2022) and (b) Ours.

In Figure 5, we use Grad-CAM (Selvaraju et al., 2017) to visualize image regions where the model focuses on for the final verdict. We observe that our model generally focuses on multiple class-discriminative features, giving the benefits of more robust and generalizable recognition performance.

F.2. t-SNE

Overall t-SNE. Figure 6 exhibits a visualization of the embedding space using t-SNE (Van der Maaten & Hinton, 2008). The t-SNE visualizes embeddings in a lower-dimensional space by aligning KL divergence with pairwise similarities in the latent space. To differentiate between target domains, we employ distinct marker styles, and for class separation, we employ different colors.

Figure 6 exhibits a visualization of the embedding space using t-SNE (Van der Maaten & Hinton, 2008). The t-SNE visualizes embeddings in a lower-dimensional space by aligning KL divergence with pairwise similarities in the latent space. To differentiate between target domains, we employ distinct marker styles, and for class separation, we employ different colors.

An ideal generalizable model would demonstrate that visual features belonging to the same class regardless domains are gathered together. This would indicate that the model relies on more domain-invariant features. From this perspective, both GVRT and our method clearly outperform ERM. ERM exhibits scattered points per domain, while both GVRT and our method display better clustering of features from the same class but different domains. Notably, in the case of our method, features on the paint domain are closely located within the red boxes, suggesting a potential degradation in generalization performance in that particular domain.

In addition, Figure 6 (d) presents box-plots for GVRT and our method, highlighting a significant observation. Our model demonstrates lower inter-domain distances between instances of the same class compared to GVRT.

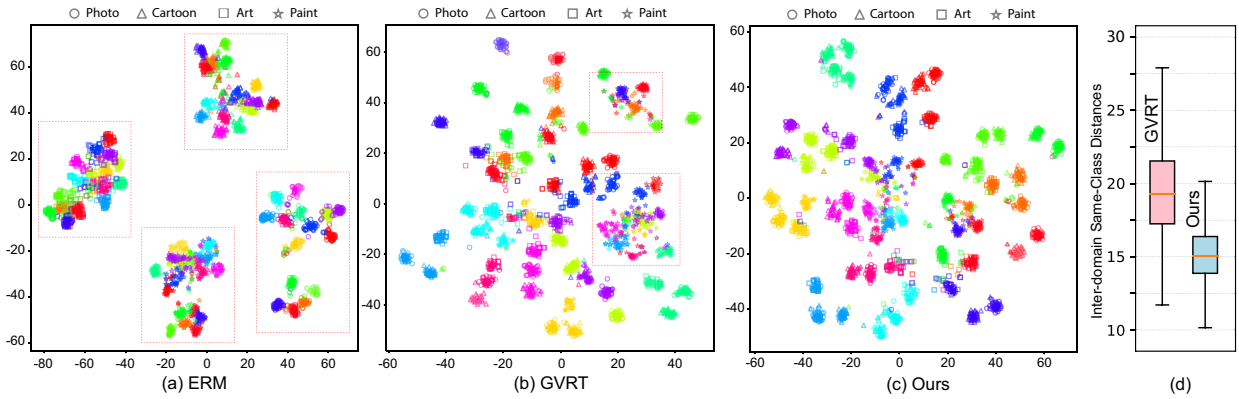


Figure 6. Visualizations by t-SNE for (a) ERM (Vapnik, 1999), (b) GVRT (Min et al., 2022), and (c) Ours. (d) We also compare inter-domain same-class distances.

Matched t-SNE. In Figure 7, we provide a detailed t-SNE (Van der Maaten & Hinton, 2008) visualization of GVRT (Min et al., 2022) and ours with matched image samples. Note that we mark diffrent shapes to represent target domains and different colors to represent classes.

In Figure 7 (a), images that belong to the same domain (ie. paint style) but diffrent classes are gathered together in the GVRT feature space. Examining the corresponding images, they have their own class discriminative characteristics like the color of beak and pattern of feather, except that they share a common domain style. In other words, the features of images can be located far away if the class discriminative characteristic is captured. Therefore, it can be inferred that the GVRT model relies more on the domain-specific features rather than domain-invaraint features for the images, limiting the ability of generalization.

Figure 7 (b) shows the distribution of images that belong to the same class but different domains. In our model, the features of same classes are located close each other unlike GVRT where the features of paint domain are located far away. In fact, our inter-domain distance is lower than GVRT. Thus, we can infer that ours captures more domain-invaraint features than GVRT for the images.

G. Complete Table Versions of Table 1, Table 2

Table 6 is the complete version of tha Table 1 in paper and Table 7 is the complete version of the Table 2.

H. Per-domain Results on Domainbed

In Table 8–11, we report per-domain results on each of the four multi-domain datasets from the large-scale DomainBed (Gulrajani & Lopez-Paz, 2020) experiments. We provide the averaged results from three independent trials. In each of the three trials, all choices, such as the dataset split, hyperparameter search, and weight initialization are selected randomly. For model selection, we used the validation set from the source domains. The reported numbers for SelfReg (Kim et al., 2021), and mDSI (Bui et al., 2021) were obtained from their respective papers, and the numbers for the remaining results were reported in the Domainbed (Gulrajani & Lopez-Paz, 2020). Note that GVRT (Min et al., 2022) and ours use multi-modal inputs (images and texts), while others only use images.

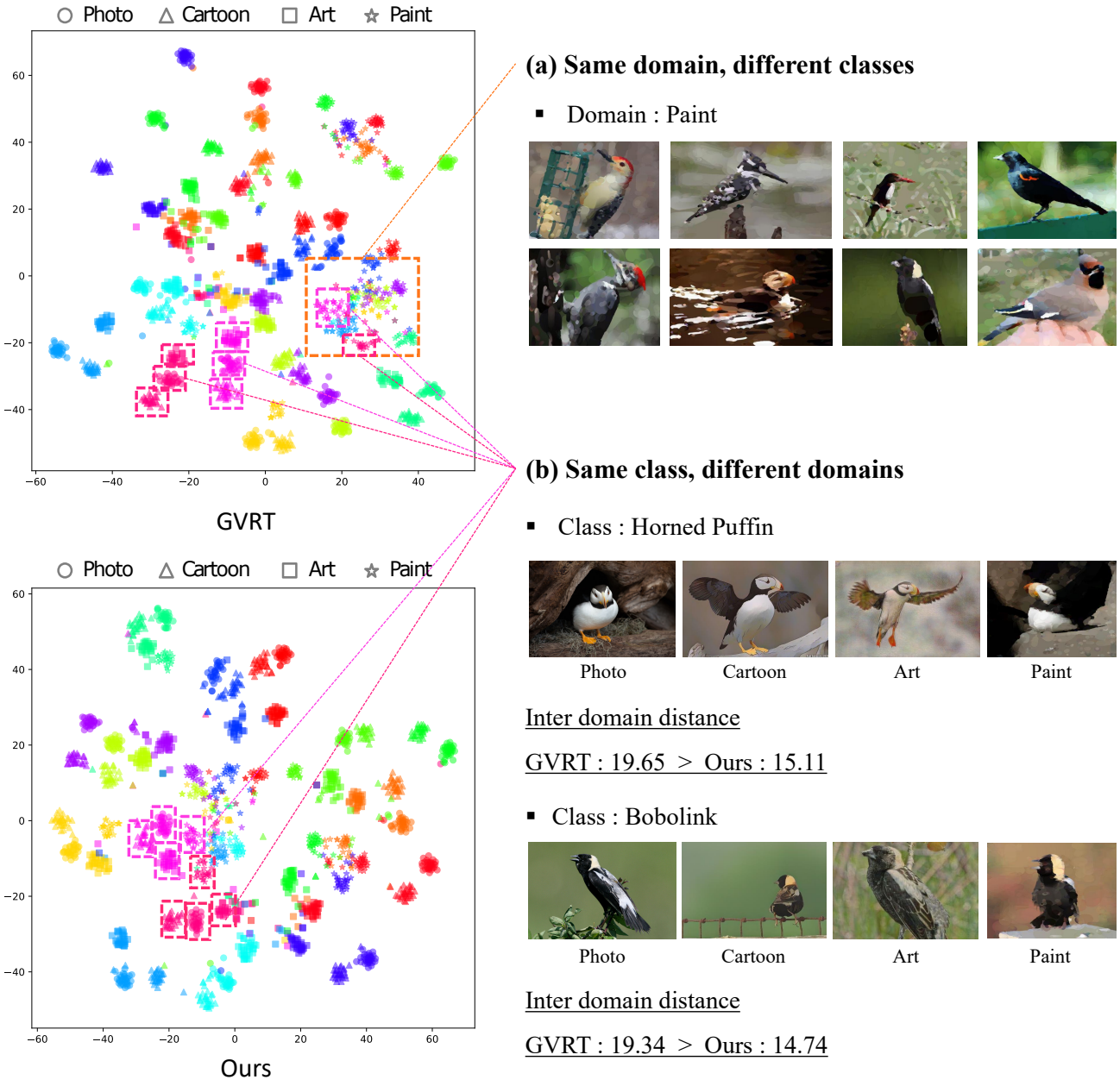


Figure 7. Visualization by t-SNE for GVRT and Ours with matched image samples.

Table 6. We assess the accuracies of out-of-distribution samples (in %) on the CUB-DG dataset using a multi-source DG task setup. In this setup, one domain serves as the test domain while the others act as source domains. We compare our approach with 13 existing DG methods. Note that our approach, as well as GVRT (Min et al., 2022), utilize multi-modal inputs (images and texts), whereas the other methods solely rely on images. *Abbr.* I: Image, T: Text.

Algorithm	Modality	Target Domain				Avg.
		Photo	Cartoon	Art	Paint	
Ours	I+T	**75.4**	**65.5**	**54.0**	**41.4**	**59.1**
GVRT (PTE) (Min et al., 2022)	I+T	74.6	64.2	52.2	37.0	57.0
CORAL (Sun & Saenko, 2016)	I	72.2	63.5	50.3	35.8	55.4
SD (Pezeshki et al., 2020)	I	71.3	62.2	50.8	34.8	54.7
SagNet (Nam et al., 2021)	I	67.4	60.7	44.0	34.2	51.6
MixStyle (Zhou et al., 2020)	I	59.0	56.7	50.3	35.8	50.4
Mixup (Yan et al., 2020)	I	67.1	55.9	51.1	27.2	50.3
DANN (Ganin et al., 2016)	I	67.5	57.0	42.8	30.6	49.5
CDANN (Li et al., 2018c)	I	65.3	55.2	43.2	30.5	48.6
VREx (Krueger et al., 2020)	I	63.9	54.9	38.6	30.1	46.9
ERM (Vapnik, 1999)	I	62.5	53.2	37.4	29.0	45.5
ARM (Zhang et al., 2020)	I	62.3	51.2	38.2	28.4	45.0
GroupDRO (Sagawa et al., 2019)	I	60.9	54.8	36.5	27.0	44.8
IRM (Arjovsky et al., 2019)	I	60.6	51.6	36.5	30.3	44.8

Table 7. The test accuracies (in %) on the DomainBed dataset (VLCS (Fang et al., 2013), PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017) and TerraIncognita (Beery et al., 2018)) in the multi-source DG task setting. We compare ours with the existing 19 other DG approaches. The reported numbers for MIRO, Fish, SelfReg, and mDSI were obtained from their respective papers, and the numbers for the remaining results were reported in the Domainbed. Note that GVRT (Min et al., 2022) and ours use multi-modal inputs (images and texts), while others only use images. *Abbr.* I: Image, T: Text

Algorithm	Modality	Dataset				Avg.
		VLCS	PACS	OfficeHome	TerraIncognita	
Ours	I+T	78.3 ± 0.4	85.7 ± 0.1	70.1 ± 0.1	49.5 ± 0.9	70.9
GVRT (PTE) (Min et al., 2022)	I+T	79.0 ± 0.2	85.1 ± 0.3	70.1 ± 0.1	48.0 ± 0.2	70.6
MIRO (Cha et al., 2022)	I	79.0 ± 0.0	85.4 ± 0.4	70.5 ± 0.4	50.4 ± 1.1	71.3
mDSDI (Bui et al., 2021)	I	79.0 ± 0.3	86.2 ± 0.2	69.2 ± 0.4	48.1 ± 1.4	70.6
CORAL (Sun & Saenko, 2016)	I	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	70.3
SagNet (Nam et al., 2021)	I	77.8 ± 0.5	86.3 ± 0.2	68.1 ± 0.1	48.6 ± 1.0	70.2
SelfReg (Kim et al., 2021)	I	77.8 ± 0.9	85.6 ± 0.4	67.9 ± 0.7	47.0 ± 0.3	69.6
Mixup (Yan et al., 2020)	I	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	69.5
Fish (Shi et al., 2021)	I	77.8 ± 0.3	85.5 ± 0.3	68.6 ± 0.4	45.1 ± 1.3	69.3
MLDG (Li et al., 2018a)	I	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	69.2
VREx (Krueger et al., 2020)	I	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	69.0
ERM (Vapnik, 1999)	I	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	68.9
DANN (Ganin et al., 2016)	I	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	68.7
RSC (Huang et al., 2020)	I	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	68.6
IRM (Arjovsky et al., 2019)	I	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	68.5
MTL (Blanchard et al., 2021)	I	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	68.5
ARM (Zhang et al., 2020)	I	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	68.3
CDANN (Li et al., 2018c)	I	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	67.9
MMD (Li et al., 2018b)	I	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	67.6
GroupDRO (Sagawa et al., 2019)	I	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	67.6

Table 8. Per-domain out-of-distribution test accuracies on the PACS (Li et al., 2017) dataset. *Abbr.* I: Image, T: Text

Algorithm	Modality	Art Painting	Cartoon	Photo	Sketch	Avg.
Ours	I+T	87.1 ± 0.5	79.8 ± 0.4	97.7 ± 0.1	78.3 ± 0.7	85.7
GVRT (PTE) (Min et al., 2022)	I+T	87.9 ± 0.3	78.4 ± 1.0	98.2 ± 0.1	75.7 ± 0.4	85.1
SagNet (Nam et al., 2021)	I	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	80.0 ± 0.4	86.3
mDSDI (Bui et al., 2021)	I	87.7 ± 0.4	80.4 ± 0.7	98.1 ± 0.3	78.4 ± 1.2	86.2
CORAL (Sun & Saenko, 2016)	I	88.3 ± 0.2	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	86.2
SelfReg (Kim et al., 2021)	I	87.9 ± 1.0	79.4 ± 1.4	96.8 ± 0.7	78.3 ± 1.2	85.6
ERM (Vapnik, 1999)	I	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5
Fish (Shi et al., 2021)	I	-	-	-	-	85.5
MIRO (Cha et al., 2022)	I	-	-	-	-	85.4
RSC (Huang et al., 2020)	I	85.4 ± 0.8	79.7 ± 1.8	97.6 ± 0.3	78.2 ± 1.2	85.2
ARM (Zhang et al., 2020)	I	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
VREx (Krueger et al., 2020)	I	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
MLDG (Li et al., 2018a)	I	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
MMD (Li et al., 2018b)	I	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
MTL (Blanchard et al., 2021)	I	87.5 ± 0.8	77.1 ± 0.5	96.4 ± 0.8	77.3 ± 1.8	84.6
Mixup (Yan et al., 2020)	I	86.1 ± 0.5	78.9 ± 0.8	97.6 ± 0.1	75.8 ± 1.8	84.6
GroupDRO (Sagawa et al., 2019)	I	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
DANN (Ganin et al., 2016)	I	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
IRM (Arjovsky et al., 2019)	I	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
CDANN (Li et al., 2018c)	I	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6

Table 9. Per-domain out-of-distribution test accuracies on the OfficeHome (Venkateswara et al., 2017) dataset. *Abbr.* I: Image, T: Text

Algorithm	Modality	Art	Clipart	Product	Real-world	Avg.
Ours	I+T	66.7 ± 1.0	55.4 ± 0.4	78.2 ± 0.4	80.0 ± 0.3	70.1
GVRT (PTE) (Min et al., 2022)	I+T	66.3 ± 0.1	55.8 ± 0.4	78.2 ± 0.4	80.4 ± 0.2	70.1
MIRO (Cha et al., 2022)	I	-	-	-	-	70.5
mDSDI (Bui et al., 2021)	I	68.1 ± 0.3	52.1 ± 0.4	76.0 ± 0.2	80.4 ± 0.2	69.2
CORAL (Sun & Saenko, 2016)	I	65.3 ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.7
Fish (Shi et al., 2021)	I	-	-	-	-	68.6
Mixup (Yan et al., 2020)	I	62.4 ± 0.8	54.8 ± 0.6	76.9 ± 0.3	78.3 ± 0.2	68.1
SagNet (Nam et al., 2021)	I	63.4 ± 0.2	54.8 ± 0.4	75.8 ± 0.4	78.3 ± 0.3	68.1
SelfReg (Kim et al., 2021)	I	63.6 ± 1.4	53.1 ± 1.0	76.9 ± 0.4	78.1 ± 0.4	67.9
MLDG (Li et al., 2018a)	I	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
ERM (Vapnik, 1999)	I	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
MTL (Blanchard et al., 2021)	I	61.5 ± 0.7	52.4 ± 0.6	74.9 ± 0.4	76.8 ± 0.4	66.4
VREx (Krueger et al., 2020)	I	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
MMD (Li et al., 2018b)	I	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.3
GroupDRO (Sagawa et al., 2019)	I	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
DANN (Ganin et al., 2016)	I	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.9
CDANN (Li et al., 2018c)	I	61.5 ± 1.4	50.4 ± 2.4	74.4 ± 0.9	76.6 ± 0.8	65.8
RSC (Huang et al., 2020)	I	60.7 ± 1.4	51.4 ± 0.3	74.8 ± 1.1	75.1 ± 1.3	65.5
ARM (Zhang et al., 2020)	I	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
IRM (Arjovsky et al., 2019)	I	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3

Table 10. Per-domain out-of-distribution test accuracies on the TerraIncognita (Beery et al., 2018) dataset. *Abbr.* I: Image, T: Text

Algorithm	Modality	L100	L38	L43	L46	Avg.
Ours	I+T	56.9 ± 3.0	45.5 ± 0.7	57.7 ± 1.4	37.8 ± 0.8	49.5
GVRT (PTE) (Min et al., 2022)	I+T	53.9 ± 1.3	41.8 ± 1.2	58.2 ± 0.9	38.0 ± 0.6	48.0
MIRO (Cha et al., 2022)	I	-	-	-	-	50.4
SagNet (Nam et al., 2021)	I	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	48.6
mDSDI (Bui et al., 2021)	I	53.2 ± 3.0	43.3 ± 1.0	56.7 ± 0.5	39.2 ± 1.3	48.1
Mixup (Yan et al., 2020)	I	59.6 ± 2.0	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG (Li et al., 2018a)	I	54.2 ± 3.0	44.3 ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.7
IRM (Arjovsky et al., 2019)	I	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
CORAL (Sun & Saenko, 2016)	I	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
SelfReg (Kim et al., 2021)	I	48.8 ± 0.9	41.3 ± 1.8	57.3 ± 0.7	40.6 ± 0.9	47.0
DANN (Ganin et al., 2016)	I	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
RSC (Huang et al., 2020)	I	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	40.8 ± 0.6	46.6
VREx (Krueger et al., 2020)	I	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
ERM (Vapnik, 1999)	I	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
CDANN (Li et al., 2018c)	I	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
MTL (Blanchard et al., 2021)	I	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
ARM (Zhang et al., 2020)	I	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
Fish (Shi et al., 2021)	I	-	-	-	-	45.1
GroupDRO (Sagawa et al., 2019)	I	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
MMD (Li et al., 2018b)	I	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2

Table 11. Per-domain out-of-distribution test accuracies on the VLCS (Fang et al., 2013) dataset. *Abbr.* I: Image, T: Text

Algorithm	Modality	Caltech	LabelMe	SUN09	VOC2007	Avg.
Ours	I+T	98.3 ± 0.3	64.6 ± 0.7	73.6 ± 2.2	76.6 ± 0.8	78.3
GVRT (PTE) (Min et al., 2022)	I+T	98.8 ± 0.1	64.0 ± 0.3	75.2 ± 0.5	77.9 ± 1.0	79.0
MIRO (Cha et al., 2022)	I	-	-	-	-	79.0
mDSDI (Bui et al., 2021)	I	97.6 ± 0.1	66.5 ± 0.4	74.0 ± 0.6	77.8 ± 0.7	79.0
CORAL (Sun & Saenko, 2016)	I	98.3 ± 0.1	66.1 ± 1.2	73.4 ± 0.3	77.5 ± 1.2	78.8
DANN (Ganin et al., 2016)	I	99.0 ± 0.3	65.1 ± 1.4	73.1 ± 0.3	77.2 ± 0.6	78.6
IRM (Arjovsky et al., 2019)	I	98.6 ± 0.1	64.9 ± 0.9	73.4 ± 0.6	77.3 ± 0.9	78.5
VREx (Krueger et al., 2020)	I	98.4 ± 0.3	64.4 ± 1.4	74.1 ± 0.4	76.2 ± 1.3	78.3
SelfReg (Kim et al., 2021)	I	96.7 ± 0.4	65.2 ± 1.2	73.1 ± 1.3	76.2 ± 0.7	77.8
SagNet (Nam et al., 2021)	I	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8
Fish (Shi et al., 2021)	I	-	-	-	-	77.8
ARM (Zhang et al., 2020)	I	98.7 ± 0.2	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
MMD (Li et al., 2018b)	I	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.5
CDANN (Li et al., 2018c)	I	97.1 ± 0.3	65.1 ± 1.2	70.7 ± 0.8	77.1 ± 1.5	77.5
ERM (Vapnik, 1999)	I	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
Mixup (Yan et al., 2020)	I	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4
MTL (Blanchard et al., 2021)	I	97.8 ± 0.4	64.3 ± 0.3	71.5 ± 0.7	75.3 ± 1.7	77.2
MLDG (Li et al., 2018a)	I	97.4 ± 0.2	65.2 ± 0.7	71.0 ± 1.4	75.3 ± 1.0	77.2
RSC (Huang et al., 2020)	I	97.9 ± 0.1	62.5 ± 0.7	72.3 ± 1.2	75.6 ± 0.8	77.1
GroupDRO (Sagawa et al., 2019)	I	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7