Supplementary Material for: Grounding Visual Representations with Texts for Domain Generalization

Seonwoo Min[1], Nokyung Park[2], Siwon Kim[3], Seunghyun Park[4], and Jinkyu Kim[2]

[1] LG AI Research, South Korea
[2] Computer Science and Engineering, Korea University, South Korea
[3] Electrical and Computer Engineering, Seoul National University, South Korea
[4] Clova AI Research, NAVER Corp., South Korea
Correspondence: jinkyukim@korea.ac.kr

This supplementary material contains details of our paper which we could not provide in the main manuscript due to page limits. We provide (1) details of the CUB-DG data split procedure, (2) implementation details, (3) additional single-domain DG results, (4) additional ablation studies results, (5) analysis with Grad-CAM, and (5) detailed DomainBed experiment results.

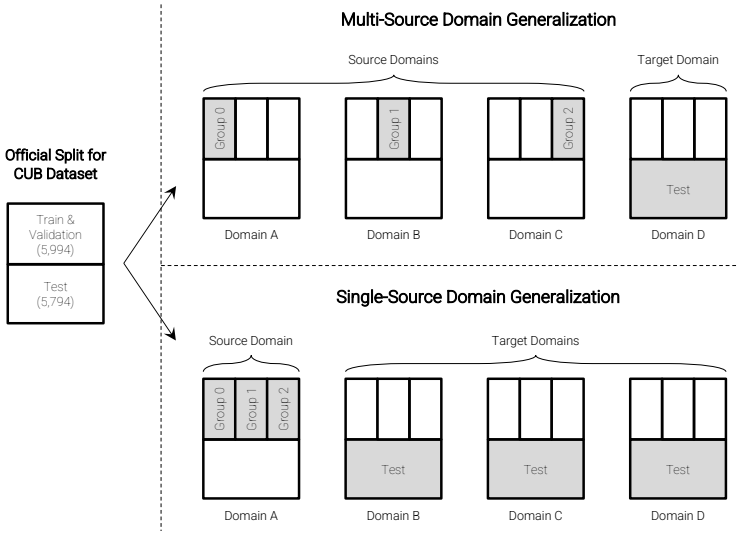

Fig. S1. Data split procedures for the CUB-DG dataset. We start from the official split of the CUB dataset. We divide the train-validation set into three disjoint groups, e.g. Group 0, Group 1, and Group 2. For the multi-source DG task, we select a different group from each source domain (gray boxes), so that the different domains do not share the siblings of the same image. For the single-source DG task, we use all three groups from a source domain.

Details of CUB-DG Data Split Procedure. In Figure S1, we show an overview of our data split procedure for the CUB-DG dataset. Note that we tried to make it close to real-life scenarios where versions of the same images do not appear in different domains.

Implementation Details. We follow the implementations of DomainBed [7], which is a unified testbed useful for evaluating DG algorithms. We use ResNet-50 [8] as the backbone of different algorithms. It is pre-trained on ImageNet [4] and produces a 2,048-dimensional latent representation from the last layer. We train each DG algorithm for 5,000 steps using Adam optimizer with a batch size of 32 for each source domain. Standard image augmentations (i.e. random cropping, horizontal flipping, color jittering, grayscale conversion, and normalization) are used during the training. For the model and training hyperparameters of each algorithm, we use the default values used in the DomainBed. In our case, we use 1.0 for λ_{align} and 1.0 for λ_{expl}. The learning rate is set to 5e−5 for the backbone parameters and 5e−4 for the newly introduced parameters.

Additional Single-Source DG Results. We provide additional results in the single-source DG task on the CUB-DG dataset. In Figure S2, we provide a heatmap that more clearly demonstrates the performance differences between ours and two baselines, i.e., ERM [24] and SD [17]. Each cell contains accuracy differences for source-target combinations, and the color blue indicates that ours performs better. Next, we provide the full results for comparing our model with six DG algorithms. Note that we excluded some algorithms (e.g. CORAL [23] and Mixup [27]). Since those algorithms explicitly match distributions across different domains, they are inapplicable for the single-source DG setting.

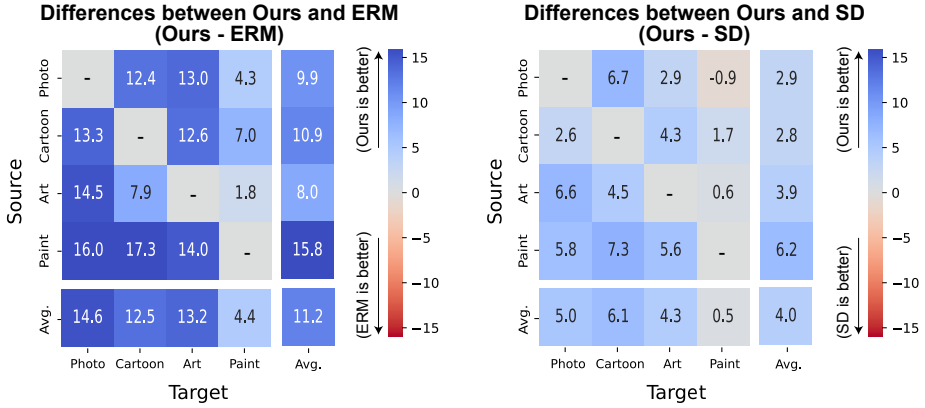

Fig. S2. Out-of-distribution test accuracies in the single-source DG setting where we train our model with a single source domain (rows) and evaluate with other remaining target domains (columns). We show performance differences between ours and two baselines (i.e. ERM [24] and SD [17]). **Blue indicates that ours is better.**

Table S1. Out-of-distribution test accuracies on the CUB-DG benchmark dataset. Here we compare six DG algorithms in the single-source DG setting. For each target domain, we report the averaged results from three models independently trained with each of the remaining domains. Note that we use the validation set (from source domains) for the model selection.

Model	Source Domain				Avg
	Photo	Cartoon	Art	Paint	
Ours w/ PTE	**69.6**	**48.1**	**41.1**	24.0	**45.7**
Ours w/ STE	69.0	**48.1**	39.2	**24.9**	45.3
SD [17]	64.6	41.9	36.9	23.6	41.7
SagNet [15]	56.0	38.1	28.7	22.2	36.3
VREX [10]	55.1	36.2	27.3	19.8	34.6
ERM [24]	55.0	35.6	27.9	19.7	34.5
ARM [28]	54.9	36.9	28.0	20.6	35.1
IRM [1]	53.1	35.6	27.6	19.3	33.9

Table S2. Results from additional ablation studies. We vary our base model in several directions and measured the performance on the multi-source DG task.

	Pre-trained Textual Encoder	Target Domain				Avg
		Photo	Cartoon	Art	Paint	
Base	CLIP [18]	74.6	**64.2**	**52.2**	**37.0**	**57.0**
(C)	MPNet [22]	74.5	63.1	49.8	37.7	56.3
	DistillBERT [20]	74.2	62.2	50.4	38.4	56.3
	MiniLM [26]	73.6	64.7	51.4	35.7	56.3

In Table S1, we report the averaged results from three models independently trained with each of the remaining domains. We observe that the proposed models outperform the others. Cross-modality supervision is especially effective in the single-source DG setting where visual representations alone deliver little information for domain invariances.

Additional Ablation Studies Results. We provide additional results from the ablation studies. We report averaged results across three independent runs in the multi-source DG setting. In Figure S3, we provide a heatmap for more extensive range of λ_{expl} (rows) and λ_{align} (columns). Again, we can see that the former

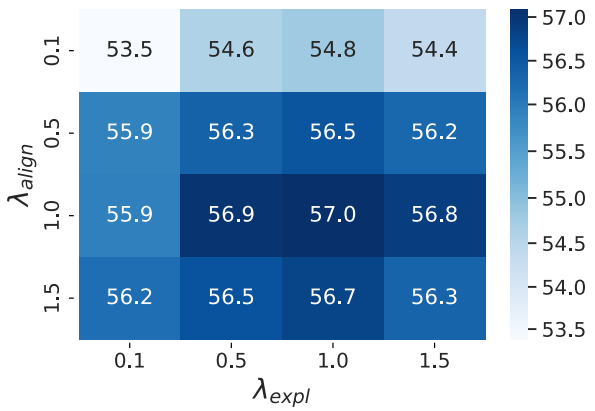

Fig. S3. Additional results from ablation studies. We report out-of-distribution test accuracies in the multi-source DG setting where we train our model with a λ_{expl} (rows) and λ_{align} (columns)

is more crucial in the training of our proposed model. In Table S2, we compare the impact of embeddings from other PTEs, i.e. MPNet [22], DistillBERT [20], and MiniLM [26]. The results show that different PTEs also successfully produce domain-invariant representations.

Analysis with Grad-CAM. As shown in Figure S4, we use Grad-CAM [21] to highlight image regions where the model attends to classify the given object. We provide two examples for different target domains (i.e. Cartoon and Photo) where we compare the model's attention maps. We observe that our proposed model captures the class-discriminative features (i.e. short pointy beak), which are compatible with the generated sentence. Note that red is the attended region.

Detailed DomainBed Experiment Results. In Table S3–S7, we provide per-domain results on each of the five multi-domain datasets from the large-scale DomainBed [7] experiments. Following their experiment protocols, we report the averaged results from three independent trials. In each trial, entire random choices (e.g. dataset splits, hyperparameter search, and weight initialization) in the study are renewed. Note that we use the validation set (from source domains) for the model selection.

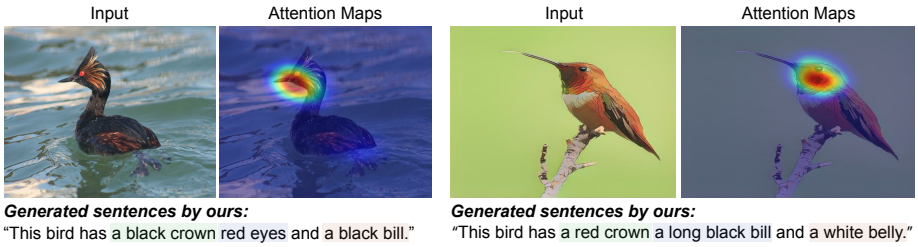

Fig. S4. We provide visualizations of attention maps (i.e. where the model sees) by Grad-CAM for ours as well as the generated sentences.

Table S3. Per-domain out-of-distribution test accuracies on the VLCS [5] dataset. The results of compared DG algorithms are excerpted from DomainBed [7]. Note that we use the train domain validation set (from source domains) for the model selection.

Method	Caltech	LabelMe	SUN09	VOC2007	Avg
Ours w/ PTE	98.8 ± 0.1	64.0 ± 0.3	**75.2 ± 0.5**	**77.9 ± 1.0**	**79.0**
CORAL [23]	98.3 ± 0.1	66.1 ± 1.2	73.4 ± 0.3	77.5 ± 1.2	78.8
DANN [6]	**99.0 ± 0.3**	**65.1 ± 1.4**	73.1 ± 0.3	77.2 ± 0.6	78.6
IRM [1]	98.6 ± 0.1	64.9 ± 0.9	73.4 ± 0.6	77.3 ± 0.9	78.5
VREx [10]	98.4 ± 0.3	64.4 ± 1.4	74.1 ± 0.4	76.2 ± 1.3	78.3
SagNet [15]	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8
ARM [28]	98.7 ± 0.2	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
ERM [24]	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
MMD [13]	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.5
CDANN [14]	97.1 ± 0.3	**65.1 ± 1.2**	70.7 ± 0.8	77.1 ± 1.5	77.5
Mixup [27]	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4
MLDG [12]	97.4 ± 0.2	65.2 ± 0.7	71.0 ± 1.4	75.3 ± 1.0	77.2
MTL [3]	97.8 ± 0.4	64.3 ± 0.3	71.5 ± 0.7	75.3 ± 1.7	77.2
RSC [9]	97.9 ± 0.1	62.5 ± 0.7	72.3 ± 1.2	75.6 ± 0.8	77.1
GroupDRO [19]	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7

Table S4. Per-domain out-of-distribution test accuracies on the PACS [11] dataset. The results of other compared DG algorithms are brought from DomainBed [7]. Note that we use the train domain validation set (from source domains) for the model selection.

Method	Art Painting	Cartoon	Photo	Sketch	Avg
Ours w/ PTE	87.9 ± 0.3	78.4 ± 1.0	**98.2 ± 0.1**	75.7 ± 0.4	85.1
SagNet [15]	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	**80.0 ± 0.4**	**86.3**
CORAL [23]	**88.3 ± 0.2**	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	86.2
ERM [24]	84.7 ± 0.4	**80.8 ± 0.6**	97.2 ± 0.3	79.3 ± 1.0	85.5
RSC [9]	85.4 ± 0.8	79.7 ± 1.8	97.6 ± 0.3	78.2 ± 1.2	85.2
ARM [28]	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
MLDG [12]	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
VREx [10]	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
Mixup [27]	86.1 ± 0.5	78.9 ± 0.8	97.6 ± 0.1	75.8 ± 1.8	84.6
MMD [13]	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
MTL [3]	87.5 ± 0.8	77.1 ± 0.5	96.4 ± 0.8	77.3 ± 1.8	84.6
GroupDRO [19]	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
DANN [6]	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
IRM [1]	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
CDANN [14]	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6

Table S5. Per-domain out-of-distribution test accuracies on the OfficeHome [25] dataset. The results of other compared DG algorithms are brought from DomainBed [7]. Note that we use the train domain validation set (from source domains) for the model selection.

Method	Art	Clipart	Product	Real-world	Avg
Ours w/ PTE	**66.3 ± 0.1**	**55.8 ± 0.4**	**78.2 ± 0.4**	**80.4 ± 0.2**	**70.1**
CORAL [23]	65.3 ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.7
Mixup [27]	62.4 ± 0.8	54.8 ± 0.6	76.9 ± 0.3	78.3 ± 0.2	68.1
SagNet [15]	63.4 ± 0.2	54.8 ± 0.4	75.8 ± 0.4	78.3 ± 0.3	68.1
MLDG [12]	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
ERM [24]	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
MTL [3]	61.5 ± 0.7	52.4 ± 0.6	74.9 ± 0.4	76.8 ± 0.4	66.4
VREx [10]	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
MMD [13]	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.3
GroupDRO [19]	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
DANN [6]	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.9
CDANN [14]	61.5 ± 1.4	50.4 ± 2.4	74.4 ± 0.9	76.6 ± 0.8	65.8
RSC [9]	60.7 ± 1.4	51.4 ± 0.3	74.8 ± 1.1	75.1 ± 1.3	65.5
ARM [28]	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
IRM [1]	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3

Table S6. Per-domain out-of-distribution test accuracies on the TerraIncognita [2] dataset. The results of other compared DG algorithms are brought from DomainBed [7]. Note that we use the train domain validation set (from source domains) for the model selection.

Method	L100	L38	L43	L46	Avg
Ours w/ PTE	53.9 ± 1.3	41.8 ± 1.2	**58.2 ± 0.9**	38.0 ± 0.6	48.0
SagNet [15]	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	**48.6**
Mixup [27]	**59.6 ± 2.0**	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG [12]	54.2 ± 3.0	**44.3 ± 1.1**	55.6 ± 0.3	36.9 ± 2.2	47.7
IRM [1]	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
CORAL [23]	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
DANN [6]	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
RSC [9]	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	**40.8 ± 0.6**	46.6
VREx [10]	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
ERM [24]	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
CDANN [14]	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
MTL [3]	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
ARM [28]	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
GroupDRO [19]	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
MMD [13]	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2

Table S7. Per-domain out-of-distribution test accuracies on the DomainNet [16] dataset. The results of other compared DG algorithms are brought from DomainBed [7]. Note that we use the train domain validation set (from source domains) for the model selection.

Method	Clip	Info	Paint	Quick	Real	Sketch	Avg
Ours w/ PTE	**62.4 ± 0.4**	**21.0 ± 0.0**	**50.5 ± 0.4**	**13.8 ± 0.3**	**64.6 ± 0.4**	**52.4 ± 0.2**	**44.1**
CORAL [23]	59.2 ± 0.1	19.7 ± 0.2	46.6 ± 0.3	13.4 ± 0.4	59.8 ± 0.2	50.1 ± 0.6	41.5
MLDG [12]	59.1 ± 0.2	19.1 ± 0.3	45.8 ± 0.7	13.4 ± 0.3	59.6 ± 0.2	50.2 ± 0.4	41.2
ERM [24]	58.1 ± 0.3	18.8 ± 0.3	46.7 ± 0.3	12.2 ± 0.4	59.6 ± 0.1	49.8 ± 0.4	40.9
MTL [3]	57.9 ± 0.5	18.5 ± 0.4	46.0 ± 0.1	12.5 ± 0.1	59.5 ± 0.3	49.2 ± 0.1	40.6
SagNet [15]	57.7 ± 0.3	19.0 ± 0.2	45.3 ± 0.3	12.7 ± 0.5	58.1 ± 0.5	48.8 ± 0.2	40.3
Mixup [27]	55.7 ± 0.3	18.5 ± 0.5	44.3 ± 0.5	12.5 ± 0.4	55.8 ± 0.3	48.2 ± 0.5	39.2
RSC [9]	55.0 ± 1.2	18.3 ± 0.5	44.4 ± 0.6	12.2 ± 0.2	55.7 ± 0.7	47.8 ± 0.9	38.9
DANN [6]	53.1 ± 0.2	18.3 ± 0.1	44.2 ± 0.7	11.8 ± 0.1	55.5 ± 0.4	46.8 ± 0.6	38.3
CDANN [14]	54.6 ± 0.4	17.3 ± 0.1	43.7 ± 0.9	12.1 ± 0.7	56.2 ± 0.4	45.9 ± 0.5	38.3
ARM [28]	49.7 ± 0.3	16.3 ± 0.5	40.9 ± 1.1	9.4 ± 0.1	53.4 ± 0.4	43.5 ± 0.4	35.5
IRM [1]	48.5 ± 2.8	15.0 ± 1.5	38.3 ± 4.3	10.9 ± 0.5	48.2 ± 5.2	42.3 ± 3.1	33.9
VREx [10]	47.3 ± 3.5	16.0 ± 1.5	35.8 ± 4.6	10.9 ± 0.3	49.6 ± 4.9	42.0 ± 3.0	33.6
GroupDRO [19]	47.2 ± 0.5	17.5 ± 0.4	33.8 ± 0.5	9.3 ± 0.3	51.6 ± 0.4	40.1 ± 0.6	33.3
MMD [13]	32.1 ± 13.3	11.0 ± 4.6	26.8 ± 11.3	8.7 ± 2.1	32.7 ± 13.8	28.9 ± 11.9	23.4

References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
2. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
3. Blanchard, G., Deshmukh, A.A., Dogan, U., Lee, G., Scott, C.: Domain generalization by marginal transfer learning. arXiv preprint arXiv:1711.07910 (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., 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)
5. Fang, C., Xu, Y., 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)
6. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
7. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
9. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
10. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Priol, R.L., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). arXiv preprint arXiv:2003.00688 (2020)
11. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
12. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
13. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5400–5409 (2018)
14. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 624–639 (2018)
15. Nam, H., et al.: Reducing domain gap by reducing style bias. In: CVPR (2021)
16. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1406–1415 (2019)
17. Pezeshki, M., Kaba, S.O., Bengio, Y., Courville, A., Precup, D., Lajoie, G.: Gradient starvation: A learning proclivity in neural networks. arXiv preprint arXiv:2011.09468 (2020)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)

19. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
20. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
22. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems **33**, 16857–16867 (2020)
23. Sun, B., 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)
24. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)
25. Venkateswara, H., Eusebio, J., Chakraborty, S., 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)
26. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers (2020)
27. Yan, S., Song, H., Li, N., Zou, L., Ren, L.: Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677 (2020)
28. Zhang, M., Marklund, H., Gupta, A., Levine, S., Finn, C.: Adaptive risk minimization: A meta-learning approach for tackling group shift. arXiv preprint arXiv:2007.02931 (2020)