Supplementary of "MAP: Towards Balanced Generalization of IID and OOD through Model-Agnostic Adapters"

Min Zhang[1], Junkun Yuan[1], Yue He[2], Wenbin Li[3], Zhengyu Chen[1], Kun Kuang[1*]
[1]Zhejiang University [2]Tsinghua University [3]Nanjing University
heyuethu@mail.tsinghua.edu.cn, liwenbin@nju.edu.cn,
{zhangmin.milab, yuanjk, chenzhengyu, kunkuang}@zju.edu.cn

A1. Supplementary Material

This supplementary material is organized as follows:

- Section A2 describes the derivation of the implicit gradient in the outer level of the bilevel optimization.
- Section A3 gives more details of six datasets including three toy and three real datasets.
- Section A4 presents additional experimental results.

A2. Derivation of Implicit Gradient

In this section, we approximate the gradient of the outer level optimization objective $\triangledown_\alpha \mathcal{L}_{ERM}(\mathcal{B}_\alpha, \omega^{(t)}, \alpha^{(t-1)})$ and for ease of notation, we omit $\mathcal{B}_\alpha$ and $\mathcal{B}_\omega$ in loss $\mathcal{L}_{ERM}$ and $\mathcal{R}$, respectively. Based on the chain rule, the gradient $\triangledown_\alpha \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)})$ can be approximated as follows:

$$
\begin{aligned}
&\triangledown_\alpha \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)}) \\
&= \triangledown_{\omega^{(t)}} \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)}) \triangledown_\alpha \omega^{(t)}(\alpha) \\
&= \triangledown_{\omega^{(t)}} \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)}) \\
&\quad \triangledown_\alpha (\omega^{(t-1)} - \eta_\omega \triangledown_\omega \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)})) \\
&= -\eta_\omega \triangledown_\omega \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)}) \\
&\quad \triangledown_\alpha \triangledown_\omega \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)}),
\end{aligned} \tag{1}
$$

where $\triangledown_\alpha$ and $\triangledown_\omega$ are partial derivatives of α and ω, respectively. η_ω is the learning rate of the model parameters ω. Motivated by [29], we make a Markov assumption that $\triangledown_\alpha \omega^{(t-1)} \approx 0$ in the last line. This assumption illustrates that given $\omega^{(t-1)}$, we do not care about how the values of α from previous steps led to $\omega^{(t-1)}$ at the t iteration step. It has already shown empirical success in previous works using the bilevel optimization (BLO) [20, 27]. For the second-order term $\triangledown_\omega \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)}) \triangledown_\alpha \triangledown_\omega \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)})$, we further propose an effective approximation by utilizing the first-order Taylor expansion of $\triangledown_\alpha \triangledown_\omega \mathcal{R}(\omega^{(t-1)})$. Specifically, for any vector $\upsilon \in \mathbb{R}^{|\omega|}$, with small $\epsilon > 0$, we have the following objective:

$$
\begin{aligned}
&\upsilon^\top \cdot \triangledown_\omega \triangledown_\alpha \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)}) \\
&\approx \frac{1}{\epsilon}(\triangledown_\alpha \mathcal{R}(\omega^{(t-1)} + \epsilon\upsilon, \alpha^{(t-1)})) \\
&\quad - \triangledown_\omega \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)}).
\end{aligned} \tag{2}
$$

Therefore $\triangledown_\alpha \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)})$ is approximated as:

$$
\begin{aligned}
&-\eta_\omega \triangledown_\omega \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)}) \\
&\quad \triangledown_\alpha \triangledown_\omega \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)}) \\
&= -\eta_\omega \frac{1}{\epsilon}(\triangledown_\alpha \mathcal{R}(\omega^{(t-1)} + \epsilon\upsilon, \alpha^{(t-1)})) \\
&\quad - \triangledown_\omega \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)}),
\end{aligned} \tag{3}
$$

where $\upsilon = \triangledown_\omega \mathcal{L}_{ERM}(\omega^{(t)}, \alpha^{(t-1)})$. The complexity of the first-order approximate is the same as OOD methods and the performance is as efficient as second-order optimization.

A3. Dataset Details

In this section, we detail describe the six datasets including three toy and three real in Figure 1. All statistics are listed in Table 1, including variant and invariant features, classes, image size, featurizer and spurious ratios of training and testing environments. These datasets are as below:

- **ColoredMNIST [3]** is a variant of the MNIST handwritten digit classification dataset [13] and is proposed by IRM [3] to evaluate the spurious correlation of the out-of-distribution (OOD) problem. The digits are colored either red or green in a way that each color is strongly correlated with a class of digits. The correlation is different during training and testing data, which leads to a spurious correlation. The correlated coefficient for two training and one testing environment is

*Corresponding author.

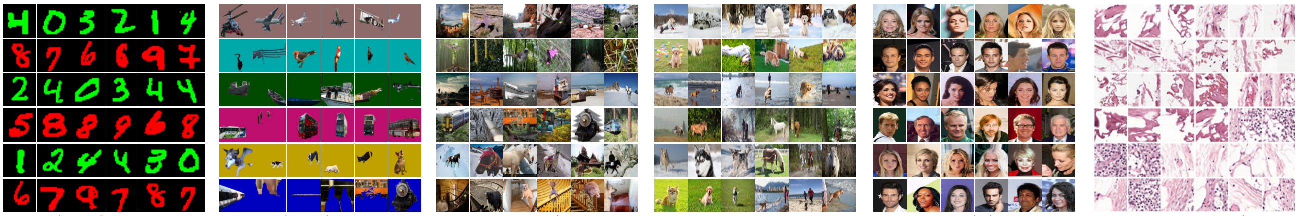

Figure 1. Examples of all datasets including ColoredMNIST, ColoredCOCO, COCOPlaces, NICO, CelebA and WILDSCamelyon.

Datasets	Variant features	Invariant features	Classes	Image size	Featurizer	Spurious ratio
ColoredMNIST [3]	color	digit	2	(2, 28, 28)	Conv4	(0.9, 0.8, 0.1)
ColoredCOCO [1]	color	object	10	(3, 64, 64)	ResNet8	(0.9, 0.8, 0.1)
COCOPlaces [1]	place	object	10	(3, 64, 64)	ResNet8	(0.9, 0.8, 0.1)
NICO [9]	background	object	2	(3, 224, 224)	ResNet18	Table 2 (left)
Celeba [18]	gender	blond	2	(3, 224, 224)	ResNet18	Table 2 (right)
WILDSCamelyon [10]	background	object	2	(3, 224, 224)	ResNet18	(1.0, 1.0, 1.0)

Table 1. Statistical information of three toy (top) and three real (bottom) by following DomainBed [8] and OoD-Bench [32].

(0.9, 0.8, 0.1), which means the occupancy of the red and the green background in class 0. Following DomainBed [8], this dataset contains 60,000 examples of dimension (2, 28, 28) and 2 classes, *i.e.*, if digits belong to the list of [0, 1, 2, 3, 4], the label is 0, otherwise, it is 1. In Table 3, we show more details of the environment split from major shifts to minor shifts.

- **ColoredCOCO [1]** is a more challenging dataset compared with ColoredMNIST. Specifically, we select ten classes from COCO [17], including airplane, bird, boat, bus, dog, horse, motorcycle, train, truck and zebra. Ten different colors are taken as spurious information like ColoredMNIST. Their RGB values are [0, 100, 0], [188, 143, 143], [255, 0, 0], [255, 215, 0], [0, 255, 0], [65, 105, 225], [0, 225, 225], [0, 0, 255], [255, 20, 147] and [160, 160, 160]. We also divide three environments, including two training environments and one testing environment. The number of samples for each training environment is 400 for each class but the testing environment is 200 for each class with the sample dimension (3, 64, 64). We use the same correlated coefficient (0.9, 0.8, 0.1), which represents that these ten biased colors are used with the corresponding coefficient and other samples use random colors.

- **COCOPlaces [1]** uses the same classes and setting (*e.g.*, image size, the number of samples or the correlated coefficient) with ColoredCOCO but different spurious information sampled from Places [34]. The spurious places as main biasd background including b/beach, c/canyon, b/building_facade, s/staircase, d/desert/sand, c/crevasse, b/bamboo_forest, f/forest/broadleaf, b/ball_pit and o/oast_house. Moreover, some random places are also used, *i.e.*, k/kasbah, l/lighthouse, p/pagoda, r/rock_arch, w/water_tower, w/waterfall, z/zen_garden.

- **NICO [9]** is a real-world dataset including photos of animals and vehicles captured in a wide range of contexts (or backgrounds). There are 10 subclasses for animals and 9 subclasses for vehicles, with each subclass having 9 or 10 different contexts. Following [32], we select a subset of this dataset to simulate the spurious correlation of different contexts and classes (animal or vehicle), which is similar to the setting of ColoredMNIST. More specifically, we make use of both classes appearing in four overlapped contexts: "on snow", "in forest", "on beach" and "on grass" to construct two training environments and one testing environment. In total, our split consists of 4,080 samples of dimension (3, 224, 224) and 2 classes of the classification task.

- **CelebA [18]** contains over 200,000 celebrity images, each of which has been annotated with 40 different attributes related to facial characteristics. It has been extensively investigated in AI fairness studies [7, 26, 25, 6] and OOD generalization [31, 32]. Following the proposed setting by GroupDRO [21], we designate "hair color" as the classification target and "gender" as the spurious attribute. We work with a subset of 27,040 images divided into three distinct environments, mimicking the ColoredMNIST setting with a significant distribution shift. To maximize the challenge of the task, we focus on the group of blond-haired males, which has the smallest number of images available.

- **WILDSCamelyon [10]** is a patch-based variant of the Camelyon17 dataset [4] curated by WILDS [10]. It comprises histopathological image slides from multiple hospitals, with data variation arising from factors

Environment	Class	NICO on snow	in forest	on beach	on grass	CelebA Class	Male	Female
Training 1	Animal	10	400	10	400	blond	462	11,671
	Vehicle	400	10	400	10	not blond	11,671	462
Training 2	Animal	20	390	20	390	blond	924	11,209
	Vehicle	390	20	390	20	not blond	11,209	924
Testing	Animal	90	10	90	10	blond	362	120
	Vehicle	10	90	10	90	not blond	120	362

Table 2. Environment splits of NICO (left) and CelebA (right) and the number of samples in each group.

		0.1		0.3		0.5		0.7		0.9	
Environment	Class	red	green	red	green	red	green	red	green	red	green
Training 1	0 (0, 1, 2, 3, 4)	10,500	1,115	10,500	1,115	10,500	1,115	10,500	1,115	10,500	1,115
	1 (5, 6, 7, 8, 9)	1,208	10,511	1,208	10,511	1,208	10,511	1,208	10,511	1,208	10,511
Training 2	0 (0, 1, 2, 3, 4)	9,306	2,308	9,306	2,308	9,306	2,308	9,306	2,308	9,306	2,308
	1 (5, 6, 7, 8, 9)	2,324	9,395	2,324	9,395	2,324	9,395	2,324	9,395	2,324	9,395
Testing	0 (0, 1, 2, 3, 4)	1,127	10,449	3,450	8,126	5,781	5,795	8,130	3,446	10,463	1,113
	1 (5, 6, 7, 8, 9)	10,449	1,180	8,219	3,538	5,924	5,833	3,559	8,198	1,191	10,566

Table 3. Environment splits of the ColoredMNIST dataset and the number of samples in each group. These ratios (*e.g.*, 0.1) represent the proportion between red and green samples in class 0 on testing data, corresponding to Table 4 of the main paper.

such as differences in patient populations, slide staining, and image acquisition. The dataset includes a total of 455,954 examples of dimension (3, 224, 224) and 2 classes, and is collected and processed by 5 hospitals.

A4. Additional Experiments

In this section, we present more experimental results based on various settings to complement the main paper.

Comparison of different structural designs of MAP. In Table 5, we analyze the impact of different connections (*i.e.*, serial or residual in Figure 4 (a) and (b)) in the main paper, different forms (*i.e.*, matrix or channel in Figure 4 (c) and (d) in the main paper) and different initializations (*i.e.*, random or eye) of IRM using the proposed MAP. In all settings, a combination of residual, matrix and random has the best performance. Other combinations also bring different performance gains, showing similar conclusions of VREx using our MAP in Table 3 in the main paper.

Could MAP perform well under different distribution shifts? In Table 4, we show the performance of all sixteen OOD methods in different distributions from major shifts to minor shifts. The performance of most OOD methods degrades as the shifts get smaller or closer to IID data, which demonstrates that these OOD methods extract invariant features while possibly losing some information that helps IID generalization. On the contrary, our MAP has good performance under different distribution shifts, which shows that MAP can learn the knowledge lost by OOD methods.

Could MAP perform well with samples of different ratios? In Table 6 without error and Table 7 with error, we generate training data and testing data with different ratios on the ColoredMNIST dataset to simulate real-world scenarios with unbalanced data distributions, *i.e.*, these ratios (*e.g.*, 0.1) represent the proportion between d_2 in training data and d_1 in testing data in Section 5.1 in the main paper. When the number of d_2 in training data is more than d_1 in testing data, especially in 0.9, the IID performance of the IID method (*i.e.*, ERM) has an increase while these OOD methods have a significant drop, which demonstrates these IID or OOD methods learn different inductive bias for IID and OOD generalizations. The proposed MAP method has a reliable and effective performance in all data ratios.

References

[1] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations, ICLR*, 2020.

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems, NeurIPS*, 34:3438–3450, 2021.

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

[4] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Methods	Major shifts → Minor shifts 0.1	0.3	0.5	0.7	0.9
ERM [28]	29.7 ± 0.6	45.5 ± 0.1 ↑	60.6 ± 0.5 ↑	85.5 ± 2.1 ↑	90.0 ± 0.2 ↑
IRM [3]	60.3 ± 3.4	53.8 ± 0.7 ↓	46.2 ± 1.2 ↓	41.7 ± 1.5 ↓	33.5 ± 0.1 ↓
VREx [12]	52.9 ± 2.9	49.6 ± 0.3 ↓	34.4 ± 1.1 ↓	22.8 ± 2.6 ↓	18.7 ± 0.1 ↓
ARM [33]	28.1 ± 0.3	43.8 ± 0.2 ↑	45.9 ± 0.4 ↑	53.3 ± 0.5 ↑	50.4 ± 0.1 ↓
GroupDRO [22]	38.5 ± 1.2	45.9 ± 0.2 ↓	48.6 ± 0.6 ↓	50.1 ± 1.3 ↓	50.9 ± 0.2 ↓
MLDG [14]	29.4 ± 0.5	40.0 ± 3.9 ↓	50.9 ± 0.1 ↓	55.0 ± 0.4 ↓	52.7 ± 0.1 ↓
MMD [15]	50.6 ± 0.1	50.3 ± 0.2 ↓	56.0 ± 0.4 ↑	53.5 ± 0.2 ↓	49.8 ± 0.1 ↓
IGA [11]	50.5 ± 0.1	45.4 ± 0.8 ↓	36.5 ± 0.1 ↓	30.0 ± 0.1 ↓	24.1 ± 0.1 ↓
SANDMask [23]	58.6 ± 6.5	53.2 ± 0.1 ↓	50.7 ± 0.3 ↓	46.5 ± 1.6 ↓	42.6 ± 0.1 ↓
Fish [24]	28.0 ± 1.2	43.3 ± 0.1 ↑	45.2 ± 0.6 ↑	42.9 ± 0.6 ↓	45.7 ± 0.1 ↑
CDANN [16]	41.7 ± 3.1	35.5 ± 0.4 ↓	29.4 ± 1.0 ↓	27.6 ± 1.5 ↓	23.1 ± 0.2 ↓
TRM [30]	44.2 ± 5.0	42.3 ± 0.1 ↓	45.7 ± 0.8 ↑	42.2 ± 0.4 ↓	31.9 ± 0.1 ↓
IB_ERM [2]	50.2 ± 0.1	50.9 ± 0.1 ↑	51.4 ± 0.1 ↑	52.4 ± 0.2 ↑	51.2 ± 0.1 ↓
IB_IRM [2]	53.8 ± 2.0	53.2 ± 0.5 ↓	48.6 ± 1.2 ↓	41.8 ± 1.7 ↓	38.1 ± 0.1 ↓
CondCAD [19]	49.2 ± 0.5	47.1 ± 0.0 ↓	36.1 ± 0.4 ↓	31.7 ± 2.7 ↓	20.9 ± 0.1 ↓
CausIRL_CORAL [5]	28.7 ± 1.3	49.5 ± 0.0 ↑	56.3 ± 0.4 ↑	49.7 ± 0.0 ↓	51.2 ± 0.1 ↑
MAP (ours)	52.6 ± 0.3	54.4 ± 0.2 ↑	62.9 ± 0.4 ↑	71.1 ± 0.7 ↑	80.5 ± 0.3 ↑

Table 4. Various distribution shifts are constructed on the ColoredMNIST dataset to simulate real-world scenarios. These ratios (*i.e.*, 0.1, 0.3, 0.5, 0.7, 0.9) represent the proportion between red and green samples in class 0 on testing data (see more details about the number of samples in Table 3). ↑ and ↓ are the increase and decrease in the model performance compared with the previous value, respectively.

Notes	Connection serial	Connection residual	Form matrix	Form channel	Init. random	Init. eye	ColoredMNIST OOD	ColoredMNIST IID	ColoredMNIST **HM**	NICO OOD	NICO IID	NICO **HM**
IRM [3]	✗	✗	✗	✗	✗	✗	60.3 ± 2.8	32.6 ± 7.0	42.3	75.8 ± 2.0	87.2 ± 0.9	81.1
+ MAP	✓	✗	✓	✗	✓	✗	41.6 ± 1.6	47.7 ± 4.3	44.4	73.6 ± 1.1	88.9 ± 0.2	80.5
	✓	✗	✓	✗	✗	✓	47.5 ± 2.1	47.6 ± 2.6	47.5	74.1 ± 1.6	88.6 ± 0.5	80.7
	✓	✗	✗	✓	✓	✗	50.3 ± 2.3	47.9 ± 1.1	49.1	75.2 ± 1.4	89.2 ± 1.2	81.6
	✓	✗	✗	✓	✗	✓	55.3 ± 1.5	49.6 ± 0.8	52.3	74.9 ± 2.3	89.1 ± 1.8	81.4
	✗	✓	✓	✗	✓	✗	57.3 ± 2.9	55.3 ± 3.2	**56.3**	76.2 ± 0.8	88.7 ± 0.4	**82.0**
	✗	✓	✓	✗	✗	✓	57.1 ± 3.5	48.0 ± 1.2	52.2	75.6 ± 2.6	88.4 ± 1.8	81.5
	✗	✓	✗	✓	✓	✗	52.1 ± 3.0	53.7 ± 0.1	52.9	74.9 ± 1.3	88.8 ± 0.6	81.3
	✗	✓	✗	✓	✗	✓	51.6 ± 0.3	54.2 ± 1.4	52.9	75.6 ± 1.1	89.0 ± 0.9	81.8

Table 5. Experiments using different forms of the adapter on ColoredMNIST and NICO. The Method in gray denotes the baseline. The specific details of connection and form are shown in Figure 4 in the main paper. Init. represents the initialization of adapter parameters.

[5] Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

[6] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

[7] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems, NeurIPS*, 34:12091–12103, 2021.

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

[9] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[10] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning, ICML*, pages 5637–5664. PMLR, 2021.

[11] Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

[12] 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, ICML*, pages 5815–5826. PMLR, 2021.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] 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, AAAI*, 2018.

Methods	0.1 OOD	0.1 IID	0.1 HM	0.3 OOD	0.3 IID	0.3 HM	0.5 OOD	0.5 IID	0.5 HM	0.7 OOD	0.7 IID	0.7 HM	0.9 OOD	0.9 IID	0.9 HM
ERM [28]	29.7	86.0	44.2	28.9	85.4	43.2	30.7	85.5	45.2	29.9	85.0	44.2	26.1	81.5	39.5
IRM [3]	60.3	32.6	42.3	61.2	33.6	43.4	64.3	31.0	41.8	60.3	33.0	42.7	43.1	51.3	46.8
VREx [12]	52.9	14.6	22.9	50.1	15.4	23.6	54.5	15.1	23.6	50.0	14.9	23.0	41.2	39.0	40.0
ARM [33]	28.1	49.9	36.0	29.6	50.7	37.4	28.2	46.3	35.1	29.9	45.8	36.2	24.0	55.1	33.4
GroupDRO [22]	38.5	51.5	44.1	34.7	50.7	41.2	34.1	50.4	40.7	38.3	50.6	43.6	33.3	57.9	42.3
MLDG [14]	29.4	50.3	34.6	29.9	50.8	37.6	30.6	50.7	38.2	32.0	50.7	39.2	32.1	50.6	39.3
MMD [15]	50.6	51.3	51.0	50.5	48.3	49.4	50.5	50.4	50.4	50.6	46.2	48.3	50.0	50.1	50.0
IGA [11]	50.5	25.0	33.4	50.8	32.0	39.3	50.7	29.7	37.5	50.3	38.9	43.9	50.5	34.9	41.3
SANDMask [23]	58.6	42.2	49.1	50.5	45.1	47.6	50.8	46.0	48.3	51.2	48.0	49.5	58.3	41.9	48.8
Fish [24]	28.0	46.4	34.9	28.2	50.1	36.1	29.0	50.3	36.8	30.9	46.1	37.0	29.8	47.6	36.7
CDANN [16]	41.7	22.6	29.3	40.2	23.6	29.7	36.8	23.5	28.7	37.5	23.6	29.0	37.2	23.3	28.7
TRM [30]	44.2	32.1	37.2	35.1	39.2	37.0	27.1	25.3	26.2	31.1	39.4	34.8	37.8	50.4	43.2
IB_ERM [2]	50.2	51.7	50.9	51.8	52.3	52.0	51.3	50.0	50.6	50.6	45.4	47.9	42.1	59.6	49.3
IB_IRM [2]	53.8	37.9	44.5	58.9	43.5	50.0	57.0	42.6	48.8	61.6	41.9	49.9	49.3	46.7	48.0
CondCAD [19]	49.2	21.1	29.5	49.5	51.9	50.7	50.6	38.2	43.5	51.7	24.0	32.8	25.3	50.4	33.7
CausIRL_CORAL [5]	28.7	50.6	36.6	30.9	50.3	38.3	42.3	56.8	48.5	32.5	67.1	43.8	25.7	63.6	36.6
MAP (ours)	52.6	71.5	**60.6**	53.1	72.0	**61.1**	52.4	71.4	**60.4**	53.8	73.5	**62.1**	54.1	72.4	**61.9**

Table 6. Different number of IID and OOD data. These ratios (*e.g.*, 0.1) represent the proportion between d_2 in training data and d_1 in testing data in Section 5.1 in the main paper.

[15] 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, CVPR*, pages 5400–5409, 2018.

[16] 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.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, ECCV*, pages 740–755. Springer, 2014.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision, ICCV*, pages 3730–3738, 2015.

[19] Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021.

[20] Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

[21] 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.

[22] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning, ICML*, pages 8346–8356. PMLR, 2020.

[23] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

[24] 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.

[25] Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *International Conference on Machine Learning, ICML*, pages 20156–20175. PMLR, 2022.

[26] Richa Singh, Puspita Majumdar, Surbhi Mittal, and Mayank Vatsa. Anatomizing bias in facial analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*, pages 12351–12358, 2022.

[27] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[28] Vladimir Vapnik and Vlamimir Vapnik. Statistical learning theory wiley. *New York*, 1(624):2, 1998.

[29] Teng Xiao, Zhengyu Chen, Donglin Wang, and Suhang Wang. Learning how to propagate messages in graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD*, pages 1894–1903, 2021.

[30] Yilun Xu and Tommi Jaakkola. Learning representations that support robust transfer of predictors. *arXiv preprint arXiv:2110.09940*, 2021.

[31] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning, ICML*, pages 25407–25437. PMLR, 2022.

[32] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7947–7958, 2022.

Methods	0.1			0.3			0.5			0.7			0.9		
	OOD	IID	HM	OOD	IID	HM	OOD	IID	HM	OOD	IID	HM	OOD	IID	HM
ERM [28]	29.7 ± 0.2	86.0 ± 0.2	44.2	28.9 ± 1.0	85.4 ± 0.1	43.2	30.7 ± 0.2	85.5 ± 0.1	45.2	29.9 ± 0.1	85.0 ± 0.1	44.2	26.1 ± 1.5	81.5 ± 0.1	39.5
IRM [3]	60.3 ± 2.8	32.6 ± 7.0	42.3	61.2 ± 2.7	33.6 ± 3.3	43.4	64.3 ± 5.0	31.0 ± 7.8	41.8	60.3 ± 5.1	33.0 ± 7.6	42.7	43.1 ± 7.0	51.3 ± 11.8	46.8
VREx [12]	52.9 ± 1.2	14.6 ± 0.3	22.9	50.1 ± 0.6	15.4 ± 0.2	23.6	54.5 ± 3.5	15.1 ± 0.1	23.6	50.0 ± 0.6	14.9 ± 0.1	23.0	41.2 ± 5.5	39.0 ± 9.8	40.0
ARM [33]	28.1 ± 0.0	49.9 ± 0.1	36.0	29.6 ± 0.1	50.7 ± 0.2	37.4	28.2 ± 0.2	46.3 ± 3.5	35.1	29.9 ± 1.4	45.8 ± 3.8	36.2	24.0 ± 0.2	55.1 ± 3.1	33.4
GroupDRO [22]	38.5 ± 1.5	51.5 ± 0.3	44.1	34.7 ± 1.5	50.7 ± 0.2	41.2	34.1 ± 2.2	50.4 ± 0.2	40.7	38.3 ± 2.8	50.6 ± 0.0	43.6	33.3 ± 2.8	57.9 ± 6.6	42.3
MLDG [14]	29.4 ± 0.6	50.3 ± 0.0	34.6	29.9 ± 0.5	50.8 ± 0.1	37.6	30.6 ± 0.7	50.7 ± 0.1	38.2	32.0 ± 0.5	50.7 ± 0.1	39.2	32.1 ± 2.2	50.6 ± 0.0	39.3
MMD [15]	50.6 ± 0.1	51.3 ± 0.6	51.0	50.5 ± 0.1	48.3 ± 1.9	49.4	50.5 ± 0.1	50.4 ± 0.1	50.4	50.6 ± 0.1	46.2 ± 3.5	48.3	50.0 ± 0.3	50.1 ± 0.2	50.0
IGA [11]	50.5 ± 0.1	25.0 ± 7.9	33.4	50.8 ± 0.1	32.0 ± 7.4	39.3	50.7 ± 0.2	29.7 ± 7.2	37.5	50.3 ± 0.1	38.9 ± 9.2	43.9	50.5 ± 0.7	34.9 ± 5.8	41.3
SANDMask [23]	58.6 ± 6.5	42.2 ± 7.2	49.1	50.5 ± 0.2	45.1 ± 4.5	47.6	50.8 ± 0.2	46.0 ± 3.6	48.3	51.2 ± 0.3	48.0 ± 2.3	49.5	58.3 ± 6.5	41.9 ± 7.2	48.8
Fish [24]	28.0 ± 1.5	46.4 ± 3.2	34.9	28.2 ± 0.5	50.1 ± 0.5	36.1	29.0 ± 0.9	50.3 ± 0.0	36.8	30.9 ± 1.2	46.1 ± 3.6	37.0	29.8 ± 0.5	47.6 ± 1.7	36.7
CDANN [16]	41.7 ± 3.5	22.6 ± 1.5	29.3	40.2 ± 4.5	23.6 ± 3.8	29.7	36.8 ± 5.3	23.5 ± 3.9	28.7	37.5 ± 5.1	23.6 ± 3.9	29.0	37.2 ± 4.5	23.3 ± 1.7	28.7
TRM [30]	44.2 ± 5.0	32.1 ± 9.5	37.2	35.1 ± 6.2	39.2 ± 5.2	37.0	27.1 ± 0.2	25.3 ± 5.5	26.2	31.1 ± 0.9	39.4 ± 10.1	34.8	37.8 ± 5.5	50.4 ± 0.1	43.2
IB_ERM [2]	50.2 ± 0.2	51.7 ± 1.7	50.9	51.8 ± 1.0	52.3 ± 1.4	52.0	51.3 ± 0.4	50.0 ± 1.9	50.6	50.6 ± 0.3	45.4 ± 2.6	47.9	42.1 ± 6.6	59.6 ± 5.2	49.3
IB_IRM [2]	53.8 ± 1.8	37.9 ± 10.0	44.5	58.9 ± 5.6	43.5 ± 6.2	50.0	57.0 ± 5.1	42.6 ± 4.7	48.8	61.6 ± 6.1	41.9 ± 8.5	49.9	49.3 ± 0.3	46.7 ± 3.0	48.0
CondCAD [19]	49.2 ± 0.5	21.1 ± 2.6	29.5	49.5 ± 2.5	51.9 ± 4.1	50.7	50.6 ± 1.2	38.2 ± 9.5	43.5	51.7 ± 1.7	24.0 ± 4.2	32.8	25.3 ± 0.8	50.4 ± 0.2	33.7
CausIRL_CORAL [5]	28.7 ± 1.3	50.6 ± 0.2	36.6	30.9 ± 0.8	50.3 ± 0.2	38.3	42.3 ± 5.3	56.8 ± 4.8	48.5	32.5 ± 13.4	67.1 ± 11.9	43.8	25.7 ± 1.4	63.6 ± 7.1	36.6
MAP (ours)	52.6 ± 0.5	71.5 ± 0.7	**60.6**	53.1 ± 1.3	72.0 ± 0.9	**61.1**	52.4 ± 1.2	71.4 ± 0.8	**60.4**	53.8 ± 2.3	73.5 ± 1.4	**62.1**	54.1 ± 1.6	72.4 ± 0.7	**61.9**

Table 7. Different number of IID and OOD data. These ratios (*e.g.*, 0.1) represent the proportion between d_2 in training data and d_1 in testing data in Section 5.1 in the main paper.

[33] 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 NeurIPS*, 34:23664–23678, 2021.

[34] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems, NeurIPS*, 27, 2014.