Supplementary Materials to
'Sharpness-Aware Gradient Matching for Domain Generalization'

The following materials are provided in this supplementary file:

- The optimal parameter settings for SAGM on each dataset.
- Full results of Table 1 in the main text.
- Robustness on ImageNet.

A. The optimal parameter settings for SAGM on each dataset

For a fair comparison, we follow the hyperparameter (HP) search protocol proposed by Cha *et al*. [6]. As mentioned in the main text, the learning rate, dropout rate, and weight decay are tuned in [1e-5, 3e-5, 5e-5], [0.0, 0.1, 0.5], and [1e-4, 1e-6] respectively. The hyperparameter α in SAGM is tuned in [1e-3, 5e-4]. To guarantee reproducibility, the optimal parameter settings for SAGM on each dataset are provided in Table A1. All experiments are conducted on a single NVIDIA A100 with Python 3.8.13, PyTorch 1.12.1, Torchvision 0.13.1 and CUDA 11.3.

Table A1. **The optimal parameter settings for SAGM on each dataset.**

Dataset	learning rate	dropout rate	weight decay	hyperparameter α
PACS	3e-5	0.5	1e-6	1e-3
VLCS	1e-5	0.5	1e-4	1e-3
OfficeHome	1e-5	0.5	1e-4	5e-4
TerraIncognita	3e-5	0.5	1e-6	5e-4
DomainNet	3e-5	0.1	1e-6	5e-4

B. Full results of Table 1 in the main text

In this section, we give the detailed results of Table 1 in the main text. Specifically, we provide the results of our SAGM and the state-of-the-art DG methods [1, 3, 4, 6–8, 10–16, 18–26] on PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet datasets in Table A2, Table A3, Table A4, Table A5 and Table A6, respectively. The results marked by †, ‡ are copied from Gulrajani and Lopez-Paz [9] and Cha *et al*. [5], respectively. Standard errors are reported from three trials, if available.

C. Robustness on ImageNet.

We use ResNet-50 as backbone and follow the standard training recipes. We use SGD optimizer with momentum of 0.9, weight decay 0.0001, base learning rate of 0.1 with linear scaling rule, batch size of 256, and total epochs of 90. The hyperparameter α is set to 0.001. The hyperparameter ρ is set to 0.05, following SAM [7].

As shown in Table A7, SAGM performs better than SAM and GSAM on ImageNet-1k and ImageNet-R. In addition, it has the same training speed as SAM and is slightly faster than GSAM.

Table A2. **Out-of-domain accuracies (%) on** `PACS` .

Algorithm	A	C	P	S	Avg
CDANN† [15]	84.6±1.8	75.5±0.9	96.8±0.3	73.5±0.6	82.6
IRM† [1]	84.8±1.3	76.4±1.1	96.7±0.6	76.1±1.0	83.5
MetaReg [2]	87.2	79.2	97.6	70.3	83.6
DANN† [8]	86.4±0.8	77.4±0.8	97.3±0.4	73.5±2.3	83.7
GroupDRO† [18]	83.5±0.9	79.1±0.6	96.7±0.3	78.3±2.0	84.4
MTL† [3]	87.5±0.8	77.1±0.5	96.4±0.8	77.3±1.8	84.6
MMD† [14]	86.1±1.4	79.4±0.9	96.6±0.2	76.5±0.5	84.7
VREx† [12]	86.0±1.6	79.1±0.6	96.9±0.5	77.7±1.7	84.9
MLDG† [13]	85.5±1.4	80.1±1.7	97.4±0.3	76.6±1.1	84.9
ARM† [24]	86.8±0.6	76.8±0.5	97.4±0.3	79.3±1.2	85.1
RSC† [10]	85.4±0.8	79.7±1.8	97.6±0.3	78.2±1.2	85.2
Mixstyle‡ [25]	86.8±0.5	79.0±1.4	96.6±0.1	78.5±2.3	85.2
ERM† [22]	84.7±0.4	80.8±0.6	97.2±0.3	79.3±1.0	85.5
CORAL† [21]	88.3±0.2	80.0±0.5	97.5±0.3	78.8±1.3	86.2
SagNet† [16]	87.4±1.0	80.7±0.6	97.1±0.1	80.0±0.4	86.3
Miro [6] (with CLIP [17])	87.4	78.2	97.2	78.7	85.4
SAM [7]	85.6±2.1	80.9±1.2	97.0±0.4	79.6±1.6	85.8
GSAM [26]	86.9±0.1	80.4±0.2	97.5±0.0	78.7±0.8	85.9
SAGM (ours)	87.4±0.2	80.2±0.3	98.0±0.2	80.8±0.6	86.6

Table A3. **Out-of-domain accuracies (%) on** `VLCS` .

Algorithm	C	L	S	V	Avg
GroupDRO† [18]	97.3±0.3	63.4±0.9	69.5±0.8	76.7±0.7	76.7
RSC† [10]	97.9±0.1	62.5±0.7	72.3±1.2	75.6±0.8	77.1
MLDG† [13]	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
ERM‡ [22]	98.0±0.3	64.7±1.2	71.4±1.2	75.2±1.6	77.3
MMD† [14]	97.7±0.1	64.0±1.1	72.8±0.2	75.3±3.3	77.5
CDANN† [15]	97.1±0.3	65.1±1.2	70.7±0.8	77.1±1.5	77.5
ARM† [24]	98.7±0.2	63.6±0.7	71.3±1.2	76.7±0.6	77.6
SagNet† [16]	97.9±0.4	64.5±0.5	71.4±1.3	77.5±0.5	77.8
Mixstyle‡ [25]	98.6±0.3	64.5±1.1	72.6±0.5	75.7±1.7	77.9
VREx† [12]	98.4±0.3	64.4±1.4	74.1±0.4	76.2±1.3	78.3
IRM† [1]	98.6±0.1	64.9±0.9	73.4±0.6	77.3±0.9	78.6
DANN† [8]	99.0±0.3	65.1±1.4	73.1±0.3	77.2±0.6	78.6
CORAL† [21]	98.3±0.1	66.1±1.2	73.4±0.3	77.5±1.2	78.8
Miro [6] (with CLIP [17])	98.3	64.7	75.3	77.8	79.0
SAM [7]	99.1±0.2	65.0±1.0	73.7±1.0	79.8±0.1	79.4
GSAM [26]	98.7±0.3	64.9±0.2	74.3±0.0	78.5±0.8	79.1
SAGM (ours)	99.0±0.2	65.2±0.4	75.1±0.3	80.7±0.8	80.0

Table A4. **Out-of-domain accuracies (%) on** `OfficeHome` .

Algorithm	A	C	P	R	Avg
Mixstyle‡ [25]	51.1±0.3	53.2±0.4	68.2±0.7	69.2±0.6	60.4
IRM† [1]	58.9±2.3	52.2±1.6	72.1±2.9	74.0±2.5	64.3
ARM† [24]	58.9±0.8	51.0±0.5	74.1±0.1	75.2±0.3	64.8
RSC† [10]	60.7±1.4	51.4±0.3	74.8±1.1	75.1±1.3	65.5
CDANN† [15]	61.5±1.4	50.4±2.4	74.4±0.9	76.6±0.8	65.7
DANN† [8]	59.9±1.3	53.0±0.3	73.6±0.7	76.9±0.5	65.9
GroupDRO† [18]	60.4±0.7	52.7±1.0	75.0±0.7	76.0±0.7	66.0
MMD† [14]	60.4±0.2	53.3±0.3	74.3±0.1	77.4±0.6	66.4
MTL† [3]	61.5±0.7	52.4±0.6	74.9±0.4	76.8±0.4	66.4
VREx† [12]	60.7±0.9	53.0±0.9	75.3±0.1	76.6±0.5	66.4
ERM† [22]	61.3±0.7	52.4±0.3	75.8±0.1	76.6±0.3	66.5
MLDG† [13]	61.5±0.9	53.2±0.6	75.0±1.2	77.5±0.4	66.8
ERM‡ [22]	63.1±0.3	51.9±0.4	77.2±0.5	78.1±0.2	67.6
SagNet† [16]	63.4±0.2	54.8±0.4	75.8±0.4	78.3±0.3	68.1
CORAL† [21]	65.3±0.4	54.4±0.5	76.5±0.1	78.4±0.5	68.7
Miro [6] (with CLIP [17])	67.5	54.6	78.0	81.6	70.5
SAM [7]	64.5±0.3	56.5±0.2	77.4±0.1	79.8±0.4	69.6
GSAM [26]	64.9±0.1	55.2±0.2	77.8±0.0	79.2 ±0.2	69.3
SAGM (ours)	65.4±0.4	57.0±0.3	78.0±0.3	80.0±0.2	70.1

Table A5. **Out-of-domain accuracies (%) on** `TerraIncognita` .

Algorithm	L100	L38	L43	L46	Avg
MMD† [14]	41.9±3.0	34.8±1.0	57.0±1.9	35.2±1.8	42.2
GroupDRO† [18]	41.2±0.7	38.6±2.1	56.7±0.9	36.4±2.1	43.2
Mixstyle‡ [25]	54.3±1.1	34.1±1.1	55.9±1.1	31.7±2.1	44.0
ARM† [24]	49.3±0.7	38.3±2.4	55.8±0.8	38.7±1.3	45.5
MTL† [3]	49.3±1.2	39.6±6.3	55.6±1.1	37.8±0.8	45.6
CDANN† [15]	47.0±1.9	41.3±4.8	54.9±1.7	39.8±2.3	45.8
ERM† [22]	49.8±4.4	42.1±1.4	56.9±1.8	35.7±3.9	46.1
VREx† [12]	48.2±4.3	41.7±1.3	56.8±0.8	38.7±3.1	46.4
RSC† [10]	50.2±2.2	39.2±1.4	56.3±1.4	40.8±0.6	46.6
DANN† [8]	51.1±3.5	40.6±0.6	57.4±0.5	37.7±1.8	46.7
IRM† [1]	54.6±1.3	39.8±1.9	56.2±1.8	39.6±0.8	47.6
CORAL† [21]	51.6±2.4	42.2±1.0	57.0±1.0	39.8±2.9	47.7
MLDG† [13]	54.2±3.0	44.3±1.1	55.6±0.3	36.9±2.2	47.8
SagNet† [16]	53.0±2.9	43.0±2.5	57.9±0.6	40.4±1.3	48.6
ERM‡ [22]	54.3±0.4	42.5±0.7	55.6±0.3	38.8±2.5	47.8
Miro [6] (with CLIP [17])	61.1	43.9	56.9	39.6	50.4
SAM [7]	46.3±1.0	38.4±2.4	54.0±1.0	34.5±0.8	43.3
GSAM [26]	50.8±0.1	39.3±0.2	59.6±0.0	38.2±0.8	47.0
SAGM (ours)	54.8±1.3	41.4±0.8	57.7±0.6	41.3±0.4	48.8

Table A6. **Out-of-domain accuracies (%) on** `DomainNet` .

Algorithm	clip	info	paint	quick	real	sketch	Avg
MMD† [14]	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
GroupDRO† [18]	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
VREx† [12]	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
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
Mixstyle‡ [25]	51.9±0.4	13.3±0.2	37.0±0.5	12.3±0.1	46.1±0.3	43.4±0.4	34.0
ARM† [24]	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
CDANN† [15]	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
DANN† [8]	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
RSC† [10]	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
SagNet† [16]	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
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
ERM† [22]	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
MLDG† [13]	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
CORAL† [21]	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
MetaReg [2]	59.8	25.6	50.2	11.5	64.6	50.1	43.6
ERM‡ [22]	62.8±0.4	20.2±0.3	50.3±0.3	13.7±0.5	63.7±0.2	52.1±0.5	43.8
Miro [6] (with CLIP [17])	63.4	21.5	50.4	12.2	65.4	52.5	44.3
SAM [7]	64.5±0.3	20.7±0.2	50.2±0.1	15.1±0.3	62.6±0.2	52.7±0.3	44.3
GSAM [26]	64.2±0.3	20.8±0.2	50.9±0.0	14.4±0.8	63.5±0.2	53.9±0.2	44.6
SAGM (ours)	64.9±0.2	21.1±0.3	51.5±0.2	14.8±0.2	64.1±0.2	53.6±0.2	45.0

Table A7. Top-1 Accuracy on ImageNet-1k and ImageNet-R and training speeds (256 images in 1 A100).

Methods	Backbone	Epoch	ImageNet-1k	ImageNet-R	Speeds
SAM	RestNet-50	90	76.9	23.8	524.65ms
GSAM	RestNet-50	90	77.2	23.6	545.37ms
SAGM	RestNet-50	90	77.4	23.9	524.65ms

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2, 3, 4

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Neural Information Processing Systems*, 31:998–1008, 2018. 2, 4

[3] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021. 1, 2, 3, 4

[4] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Neural Information Processing Systems*, 2021. 1

[5] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. 1

[6] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4

[7] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 4

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(1):2096–2030, 2016. 1, 2, 3, 4

[9] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 1

[10] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *European Conference on Computer Vision*, 2020. 1, 2, 3, 4

[11] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *International Conference on Computer Vision*, 2021. 1

[12] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020. 1, 2, 3, 4

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2, 3, 4

[14] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4

[15] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2, 3, 4

[16] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 4

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[18] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4

[19] 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. 1

[20] 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. 1

[21] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016. 1, 2, 3, 4

[22] V Vapnik. Statistical learning theory. *NY: Wiley*, 1998. 1, 2, 3, 4

[23] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, 2020. 1

[24] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020. 1, 2, 3, 4

[25] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 4

[26] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022. 1, 2, 3, 4