Supplemental Material to
SelfReg: Self-supervised Contrastive Regularization for Domain Generalization

Content

This supplementary material provides details on our experiment of hyperparameter tuning (Section 1), our detailed experiment on DomainBed benchmark (Section 2), and more examples of our analysis with Grad-CAM [10] (Section 3).

1. Hyperparameter Tuning

In Table 1 and 2, we provide the experimental results of hyperparameter tuning by grid search about scale factor of feature and logit-level regularization, respectively. To see the performance variance, we trained each model more than 5 times for each test domain and report the average image recognition accuracy and its standard deviation.

In Table 3, we provide the experimental results by the changes of the CDPL's learning rate. The CDPL is only trained by $\mathcal{L}_{\text{selfreg}}$, not classification loss. Thus, the same learning rate with the other parts of the model is may not big enough to fully optimize the CDPL. We experimented by increasing the learning rate of the CDPL, and confirmed the highest performance when the learning rate was 15 times larger than the existing learning rate. Note that we trained each model 20 times for each test domain.

2. Detailed DomainBed Experiments

Table 4-9 provide detailed results for each domain of DomainBed [5]. We report the results of two model selection methods, training-domain validation set method and test-domain validation set method. Training-domain validation set method is a classic setup of domain generalization task. Specifically, (i) split each training domain into training and validation subsets (ii) pool the validation subsets of each training domain to create an overall validation set (iii) choose the model maximizing the accuracy on the overall validation set [5]. Test-domain validation set is the method that choose the model maximizing the accuracy on a validation set that follows the distribution of the test domain [5].

In Table 4-9, SelfReg (ours)† does not include Inter-domain Curriculum Learning (IDCL) and SelfReg with stochastic weight averaging (SWA) [7] techniques, and we provide the performance of our SelfReg with SWA technique also. Since DomainBed is supposed to be evaluated every N steps, we needed to modify the codes to apply the SWA technique. We modified the code to evaluate model on the test set after completing 5000 steps learning with SWA technique. Specifically, we performed weight averaging every step from 4000 to 5000 steps. As shown in Table 4 of the main paper, SelfReg with SWA shows the state-of-the-art performance in DomainBed benchmark.

3. More GradCAM Visualizations

In Figure 1, we provide more examples for different target domains where we compare the model's attention maps using Grad-CAM [10].

References

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

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018. 6

[3] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1657–1664, 2013. 5

[4] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2551–2559, 2015. 4

[5] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 1, 4, 5, 6

[6] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 4

[7] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 1

[8] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3, 4, 5

[9] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019. 6

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 1, 4

[11] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 4, 5, 6

[12] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5018–5027, 2017. 5

Table 1. Grid search results about λ_{feature} on PACS [8].

λ_{feature}	Test Domain				Average
	Photo	*Art Painting*	*Cartoon*	*Sketch*	
1.0	95.43 ± 0.5	80.54 ± 1.3	76.02 ± 1.2	73.55 ± 2.1	81.39 ± 0.7
0.5	95.72 ± 0.4	79.50 ± 1.4	76.00 ± 1.0	74.18 ± 2.2	81.35 ± 0.6
0.4	95.57 ± 0.4	79.74 ± 1.4	76.01 ± 1.0	74.81 ± 2.0	81.53 ± 0.8
0.3	95.63 ± 0.4	79.64 ± 1.2	76.41 ± 1.7	74.68 ± 1.7	$\mathbf{81.59 \pm 0.5}$
0.2	95.64 ± 0.4	80.08 ± 1.2	75.72 ± 1.2	74.45 ± 2.2	81.47 ± 0.7
0.1	95.52 ± 0.4	80.29 ± 0.9	75.77 ± 1.1	74.36 ± 2.9	81.48 ± 0.7
0.0 (baseline)	95.66 ± 0.4	79.89 ± 1.3	75.61 ± 1.5	73.33 ± 2.8	81.12 ± 0.8

Table 2. Grid search results about λ_{logit} on PACS [8].

λ_{logit}	Test Domain				Average
	Photo	*Art Painting*	*Cartoon*	*Sketch*	
1.0	96.19 ± 0.3	81.59 ± 1.2	76.98 ± 1.3	75.71 ± 1.3	$\mathbf{82.62 \pm 0.5}$
0.9	96.03 ± 0.3	81.66 ± 0.4	77.35 ± 1.0	75.07 ± 1.0	82.53 ± 0.4
0.8	96.13 ± 0.5	82.00 ± 1.5	77.47 ± 0.8	74.18 ± 2.1	82.45 ± 0.6
0.5	96.03 ± 0.7	81.46 ± 0.9	77.24 ± 1.8	75.08 ± 1.5	82.45 ± 0.4
0.1	95.67 ± 0.4	81.09 ± 0.8	77.26 ± 0.6	74.62 ± 1.5	82.16 ± 0.3
0.0 (baseline)	95.66 ± 0.4	79.89 ± 1.3	75.61 ± 1.5	73.33 ± 2.8	81.12 ± 0.8

Table 3. Experimental results about the learning rate of the CDPL on PACS [8]. The learning rates of the other parts are 0.004.

Learning rate	Test Domain				Average
	Photo	*Art Painting*	*Cartoon*	*Sketch*	
A. 0 (untrained)	96.19 ± 0.2	82.68 ± 0.9	78.53 ± 0.5	77.11 ± 0.9	83.63 ± 0.4
B. 0.004	96.22 ± 0.3	82.34 ± 0.5	78.43 ± 0.7	77.47 ± 0.8	83.62 ± 0.3
C. B$\times$2	96.29 ± 0.2	82.14 ± 0.8	78.84 ± 0.8	77.22 ± 1.1	83.63 ± 0.4
D. B$\times$5	96.04 ± 0.2	82.17 ± 0.8	79.00 ± 0.8	77.62 ± 1.0	83.71 ± 0.4
E. B$\times$10	96.21 ± 0.3	82.51 ± 0.6	78.96 ± 0.6	77.35 ± 0.8	83.76 ± 0.3
F. B$\times$15	96.21 ± 0.3	82.41 ± 0.8	79.12 ± 0.5	77.58 ± 0.7	$\mathbf{83.83 \pm 0.3}$
G. B$\times$20	96.16 ± 0.3	82.15 ± 0.7	78.98 ± 0.7	77.43 ± 0.6	83.68 ± 0.3

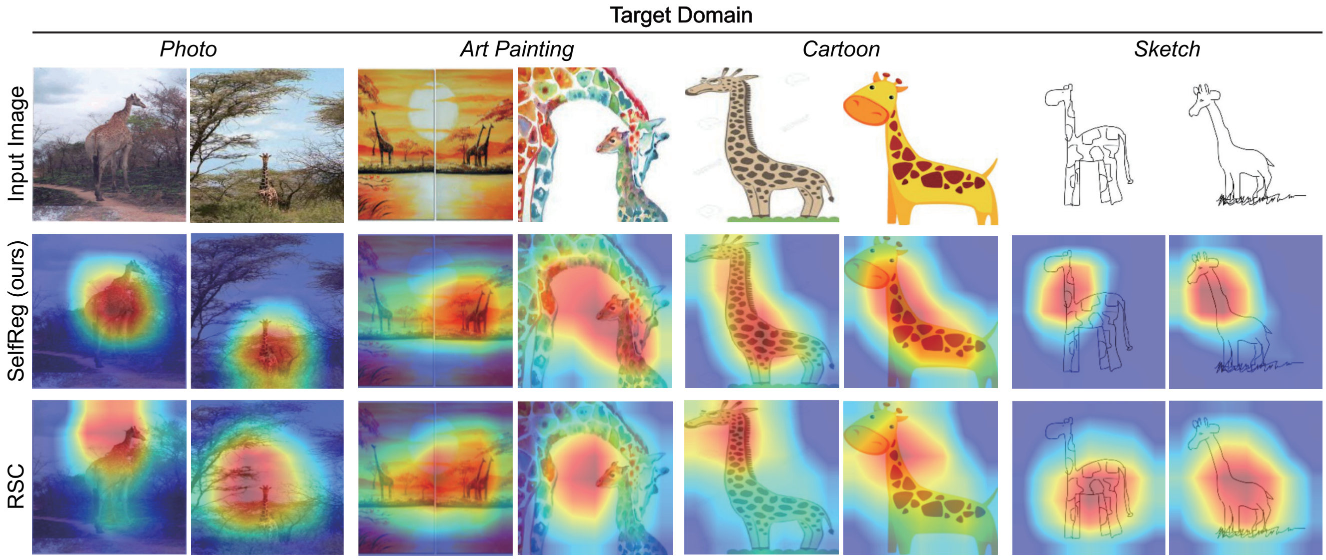

Figure 1. Original images with a giraffe for different domains (1st row). We provide visualizations of Grad-CAM [10] for ours and RSC [6], which localizes class-discriminative regions. *Data*: PACS [8]

Table 4. Detailed scores on ColoredMNIST [1] in DomainBed [5].

Model selection: training-domain validation set				
Algorithm	+90%	+80%	-90%	Avg
SelfReg (ours)†	72.2 ± 0.5	73.7 ± 0.2	10.5 ± 0.3	52.1 ± 0.2
SelfReg w/ SWA (ours)	71.3 ± 0.4	73.0 ± 0.3	10.6 ± 0.2	51.6 ± 0.2
ERM [11]	71.7 ± 0.1	72.9 ± 0.2	10.0 ± 0.1	51.5
Model selection: test-domain validation set (oracle)				
Algorithm	+90%	+80%	-90%	Avg
SelfReg (ours)†	71.3 ± 0.4	73.4 ± 0.2	29.3 ± 2.1	58.0 ± 0.7
SelfReg w/ SWA (ours)	71.7 ± 0.1	72.8 ± 0.2	28.4 ± 1.4	57.7 ± 0.4
ERM [11]	71.8 ± 0.4	72.9 ± 0.1	28.7 ± 0.5	57.8

Table 5. Detailed scores on RotatedMNIST [4] in DomainBed [5].

Model selection: training-domain validation set							
Algorithm	0	15	30	45	60	75	Avg
SelfReg w/ SWA (ours)	95.8 ± 0.2	98.8 ± 0.1	98.9 ± 0.1	99.0 ± 0.0	99.0 ± 0.0	96.8 ± 0.1	98.0 ± 0.1
SelfReg (ours)†	95.7 ± 0.3	99.0 ± 0.1	98.9 ± 0.1	99.0 ± 0.1	98.9 ± 0.1	96.6 ± 0.1	98.0 ± 0.2
ERM [11]	95.9 ± 0.1	98.9 ± 0.0	98.8 ± 0.0	98.9 ± 0.0	98.9 ± 0.0	96.4 ± 0.0	98.0
Model selection: test-domain validation set (oracle)							
Algorithm	0	15	30	45	60	75	Avg
SelfReg w/ SWA (ours)	96.1 ± 0.3	98.8 ± 0.0	99.0 ± 0.0	98.9 ± 0.1	99.0 ± 0.0	96.2 ± 0.2	98.0 ± 0.0
SelfReg (ours)†	96.0 ± 0.3	98.9 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	96.8 ± 0.1	98.1 ± 0.1
ERM [11]	95.3 ± 0.2	98.7 ± 0.1	98.9 ± 0.1	98.7 ± 0.2	98.9 ± 0.0	96.2 ± 0.2	97.8

Table 6. Detailed scores on VLCS [3] in DomainBed [5].

Algorithm	C	L	S	V	Avg
Model selection: training-domain validation set					
SelfReg (ours)†	96.7 ± 0.4	65.2 ± 1.2	73.1 ± 1.3	76.2 ± 0.7	77.8 ± 0.9
SelfReg w/ SWA (ours)	97.4 ± 0.4	63.5 ± 0.3	72.6 ± 0.1	76.7 ± 0.7	77.5 ± 0.0
ERM [11]	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
Model selection: test-domain validation set (oracle)					
SelfReg (ours)†	97.9 ± 0.4	66.7 ± 0.1	73.5 ± 0.7	74.7 ± 0.7	78.2 ± 0.1
SelfReg w/ SWA (ours)	98.2 ± 0.3	63.9 ± 0.8	72.2 ± 0.1	75.5 ± 0.4	77.5 ± 0.2
ERM [11]	97.6 ± 0.3	67.9 ± 0.7	70.9 ± 0.2	74.0 ± 0.6	77.6

Table 7. Detailed scores on PACS [8] in DomainBed [5].

Algorithm	A	C	P	S	Avg
Model selection: training-domain validation set					
SelfReg w/ SWA (ours)	85.9 ± 0.6	81.9 ± 0.4	96.8 ± 0.1	81.4 ± 0.6	86.5 ± 0.3
SelfReg (ours)†	87.9 ± 1.0	79.4 ± 1.4	96.8 ± 0.7	78.3 ± 1.2	85.6 ± 0.4
ERM [11]	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5
Model selection: test-domain validation set (oracle)					
SelfReg w/ SWA (ours)	87.5 ± 0.1	83.0 ± 0.1	97.6 ± 0.1	82.8 ± 0.2	87.7 ± 0.1
SelfReg (ours)†	87.9 ± 0.5	80.6 ± 1.1	97.1 ± 0.4	81.1 ± 1.3	86.7 ± 0.8
ERM [11]	86.5 ± 1.0	81.3 ± 0.6	96.2 ± 0.3	82.7 ± 1.1	86.7

Table 8. Detailed scores on OfficeHome [12] in DomainBed [5].

Algorithm	A	C	P	R	Avg
Model selection: training-domain validation set					
SelfReg w/ SWA (ours)	64.9 ± 0.8	55.4 ± 0.6	78.4 ± 0.2	78.8 ± 0.1	69.4 ± 0.2
SelfReg (ours)†	63.6 ± 1.4	53.1 ± 1.0	76.9 ± 0.4	78.1 ± 0.4	67.9 ± 0.7
ERM [11]	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
Model selection: test-domain validation set (oracle)					
SelfReg w/ SWA (ours)	66.1 ± 0.4	56.8 ± 0.3	78.5 ± 0.2	79.9 ± 0.1	70.3 ± 0.1
SelfReg (ours)†	64.2 ± 0.6	53.6 ± 0.7	76.7 ± 0.3	77.9 ± 0.5	68.1 ± 0.3
ERM [11]	61.7 ± 0.7	53.4 ± 0.3	74.1 ± 0.4	76.2 ± 0.6	66.4

Table 9. Detailed scores on TerraIncognita [2] in DomainBed [5].

Model selection: training-domain validation set					
Algorithm	L100	L38	L43	L46	Avg
SelfReg w/ SWA (ours)	56.8 ± 0.9	44.7 ± 0.6	59.6 ± 0.3	42.9 ± 0.8	51.0 ± 0.4
SelfReg (ours)†	48.8 ± 0.9	41.3 ± 1.8	57.3 ± 0.7	40.6 ± 0.9	47.0 ± 0.3
ERM [11]	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
Model selection: test-domain validation set (oracle)					
Algorithm	L100	L38	L43	L46	Avg
SelfReg w/ SWA (ours)	60.6 ± 0.9	47.3 ± 0.4	60.5 ± 0.7	44.6 ± 1.1	53.2 ± 0.3
SelfReg (ours)†	60.0 ± 2.3	48.8 ± 1.0	58.6 ± 0.8	44.0 ± 0.6	52.8 ± 0.9
ERM [11]	59.4 ± 0.9	49.3 ± 0.6	60.1 ± 1.1	43.2 ± 0.5	53.0

Table 10. Detailed scores on DomainNet [9] in DomainBed [5].

Model selection: training-domain validation set							
Algorithm	Clip	Info	Paint	Quick	Real	Sketch	Avg
SelfReg w/ SWA (ours)	62.4 ± 0.1	22.6 ± 0.1	51.8 ± 0.1	14.3 ± 0.1	62.5 ± 0.2	53.8 ± 0.3	44.6 ± 0.1
SelfReg (ours)†	58.5 ± 0.1	20.7 ± 0.1	47.3 ± 0.3	13.1 ± 0.3	58.2 ± 0.2	51.1 ± 0.3	41.5 ± 0.2
ERM [11]	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
Model selection: test-domain validation set (oracle)							
Algorithm	Clip	Info	Paint	Quick	Real	Sketch	Avg
SelfReg w/ SWA (ours)	62.4 ± 0.1	22.5 ± 0.2	51.7 ± 0.1	14.7 ± 0.2	62.7 ± 0.1	53.7 ± 0.4	44.6 ± 0.1
SelfReg (ours)†	58.5 ± 0.1	20.7 ± 0.1	48.0 ± 0.2	13.2 ± 0.3	58.2 ± 0.2	51.1 ± 0.3	41.6 ± 0.1
ERM [11]	58.6 ± 0.3	19.2 ± 0.2	47.0 ± 0.3	13.2 ± 0.2	59.9 ± 0.3	49.8 ± 0.4	41.3