Supplementary Materials for
Confidence Calibration for Domain Generalization under Covariate Shift

Yunye Gong[1], Xiao Lin[1], Yi Yao[1], Thomas G. Dietterich[2], Ajay Divakaran[1], and Melinda Gervasio[1]
[1]SRI International, [2]School of Electrical Engineering and Computer Science, Oregon State University
[1]first.last@sri.com,[2]tgd@oregonstate.edu

S.1. Calibration of improved classifiers

In this section, we provide additional results and discussion on applying proposed calibration algorithms on improved classifiers trained to maintain accuracy across domains.

S.1.1. Domain generalization accuracy

In the main paper we report experiments with classifiers not specifically optimized with respect to domain generalization accuracy, as described in Section 5.1. Table S.1 lists corresponding classification accuracy on Office-Home [4]. As we assume that no target data is available at the training and calibration stage, the classifier is not adapted to the target domain. As a result, classification accuracy is relatively low compared to the reported performance of TransCal [6] where unlabeled target data is used for unsupervised domain adaptation.

Real	Product	Art	Clipart	**Average**
62.30	60.14	45.78	37.97	51.55

Table S.1: Domain generalization accuracy (%) on Office-Home [4]

S.1.2. multi-source classifiers

In addition to experimental set-ups described in Section 5.1 of the main paper, we study the alternative set-up considering multiple source domains. For experiments on Office-Home [4], each classifier is trained and calibrated on 3 domains and tested on the holdout target domain. The classification accuracy and ECE scores are listed in Table S.2. Comparing Table S.1 and S.2, we see that classifiers trained on multiple sources produce better generalization accuracy. They also have slightly lower ECE on Real and Product, which shows that the calibration performance of classifiers trained using multiple sources can, to some extent, generalize to new domains, especially for domains that are similar to their training data (ResNet is pre-trained on ImageNet). However, for domains that are more different (Art, Clipart), applying the proposed calibration methods to classifiers trained using a single domain produces lower ECE.

	Real	Product	Art	Clipart	**Average**
Accuracy	75.55	73.66	59.17	45.03	63.35
ECE	3.35	2.80	5.64	16.01	6.95

Table S.2: Calibration of multi-source classifiers on Office-Home [4]

S.1.3. classifiers with domain-invariant features

We further study the effect of applying proposed calibration algorithms on top of domain-invariant features learned via domain generalization [1]. To better manifest the benefit from both domain-invariant features and calibration, we collect a dataset with 10 domains from simulations in the game environment StarCraft2 [5]. For each domain, we consider a different StarCraft2 map and collect data over 6 different unit formation classes. For each experiment we use 4 domains for training the classifiers, 4 domains for calibrating the classifiers and 2 holdout domains for testing. We learn MLP classifiers on top of ResNet18 pretrained on ILSVRC-1000. Table S.3 compares the calibration performance with and without domain-invariant features. It is shown that with domain-invariant features, ECE is further reduced, which indicates that calibration and feature alignment can be complementary.

We also note that, in general, it is a design choice on how to distribute domains between training and calibration. In this paper, we focus on one end of the trade-off, i.e., calibration.

ECE (%)	Uncalibrated	Target-Only	Ours
w/o	23.36	5.17	6.30
w	11.63	2.96	4.41

Table S.3: Performance comparison on StarCraft2. Four domains for training, four for calibration, and two as the target domains.

Source → Target	uncalibrated	source-only	target-only (oracle)	**Set-level**	**Cluster -level:NN**	**Cluster-level Regression**
I, S→ Q	21.33±0.20	19.00±0.20	0.28±0.13	10.81±0.17*	**10.60**±0.17*	19.88±0.20
I, C→ Q	27.12±0.19	23.46±0.18	0.53±0.13	11.08±0.14*	**6.46**±0.14*	9.21±0.14*
I, P→ Q	26.49±0.17	23.22±0.16	0.99±0.10	14.35±0.13*	**8.03**±0.11*	8.79±0.11*
I, R→ Q	23.41±0.18	23.67±0.19	0.55±0.12	12.34±0.14*	**6.24**±0.12*	7.68±0.13*
S, C→ Q	19.56±0.21	19.32±0.21	0.97±0.17	**6.92**±0.18*	8.78±0.17*	16.98±0.19*
S, P→ Q	20.95±0.19	18.46±0.18	0.34±0.14	**10.36**±0.16*	11.47±0.16*	22.86±0.20
S, R→ Q	15.59±0.18	15.66±0.18	0.81±0.15	**7.17**±0.16*	10.22±0.17*	12.29±0.20*
C, P→ Q	19.56±0.16	18.03±0.16	0.69±0.14	9.60±0.14*	**5.90**±0.14*	8.50±0.15*
C, R→ Q	18.77±0.17	19.54±0.17	0.67±0.14	7.48±0.14*	6.74±0.14*	**5.87**±0.14*
P, R→ Q	21.37±0.16	22.01±0.17	0.96±0.11	9.95±0.12*	6.57±0.12*	**5.15**±0.11*
Q, S→ I	28.73±0.15	28.73±0.15	2.20±0.12	10.73±0.13*	**8.07**±0.13*	9.71±0.14*
Q, C→ I	23.96±0.16	26.94±0.17	1.41±0.14	8.95±0.14*	**6.45**±0.14*	7.46±0.14*
Q, P→ I	20.07±0.14	21.77±0.15	1.93±0.13	8.34±0.16*	**5.05**±0.13*	5.61±0.14*
Q, R→ I	25.94±0.17	30.28±0.18	1.40±0.14	11.98±0.16*	6.39±0.15*	**3.50**±0.14*
S, C→ I	32.63±0.19	32.33±0.19	1.82±0.15	**11.55**±0.16*	13.14±0.16*	12.10±0.17*
S, P→ I	23.01±0.17	20.15±0.17	1.82±0.14	**6.79**±0.15*	9.62±0.15*	12.55±0.16*
S, R→ I	19.36±0.16	19.45±0.16	2.35±0.15	**3.97**±0.15*	9.72±0.16*	27.63±0.22
C, P→ I	20.87±0.18	19.12±0.16	1.85±0.14	**5.61**±0.16*	6.97±0.16*	10.92±0.17*
C, R→ I	23.88±0.18	24.82±0.18	1.50±0.13	**4.07**±0.16*	7.39±0.17*	9.78±0.18*
P, R→ I	21.48±0.18	22.24±0.18	1.85±0.14	**1.93**±0.15*	6.31±0.16*	20.05±0.22*
Q, I→ S	15.21±0.22	17.13±0.22	0.85±0.16	4.33±0.21*	2.65±0.20*	**2.06**±0.19*
Q, C→ S	17.66±0.24	20.56±0.24	1.96±0.22	1.63±0.21*	**1.04**±0.19*	1.28±0.20*
Q, P→ S	18.57±0.23	20.39±0.24	1.35±0.20	5.38±0.22*	3.32±0.23*	**3.10**±0.22*
Q, R→ S	15.03±0.23	19.03±0.23	2.78±0.21	1.45±0.20*	**1.31**±0.19*	2.72±0.22*
I, C→ S	20.89±0.25	17.43±0.25	1.47±0.22	**1.14**±0.20*	3.10±0.24*	6.78±0.25*
I, P→ S	18.03±0.24	14.83±0.24	1.48±0.22	**0.93**±0.17*	4.75±0.22*	6.67±0.24*
I, R→ S	16.33±0.23	16.59±0.23	2.32±0.22	**3.27**±0.23*	4.14±0.23*	16.69±0.28
C, P→ S	16.47±0.25	14.59±0.25	2.09±0.23	**1.43**±0.21*	3.31±0.20*	7.84±0.25*
C, R→ S	14.07±0.24	14.94±0.24	3.97±0.23	7.23±0.23*	**4.97**±0.24*	10.12±0.27*
P, R→ S	14.27±0.23	14.98±0.23	1.83±0.22	8.36±0.23*	**1.82**±0.20*	15.65±0.29

Table S.4: Quantitative evaluation of calibration on the DomainNet dataset [3] for experiments using Quickdraw (Q), Infograph (I) or Sketch (S) as the target domain. For each experiment, we evaluate each of the three proposed algorithms over 1000 experiments with randomly selected test data from the target domain. Results with statistically significant improvement against source-only method are highlighted with asterisks.

Source → Target	uncalibrated	source-only	target-only (oracle)	**Set-level**	**Cluster -level:NN**	**Cluster-level Regression**
Q, I→ C	12.09±0.17	14.05±0.17	1.83±0.16	**1.00**±0.14*	3.23±0.16*	4.64±0.17*
Q, S→ C	17.18±0.18	17.18±0.18	2.74±0.17	**2.45**±0.17*	3.83±0.18*	3.71±0.17*
Q, P→ C	14.33±0.18	16.09±0.18	2.08±0.18	**1.07**±0.14*	4.03±0.18*	5.08±0.18*
Q, R→ C	10.05±0.19	13.98±0.19	1.84±0.17	5.92±0.19*	7.82±0.19*	**4.52**±0.19*
I, S→ C	11.69±0.20	9.21±0.20	3.32±0.19	6.01±0.19*	**3.94**±0.19*	7.51±0.20*
I, P→ C	12.75±0.18	9.61±0.18	2.75±0.18	5.59±0.18*	**4.13**±0.18*	9.59±0.20
I, R→ C	10.26±0.19	10.52±0.19	1.94±0.17	11.21±0.19	**8.38**±0.19*	14.84±0.21
S, P→ C	11.49±0.19	8.80±0.19	2.88±0.19	6.77±0.19*	**4.41**±0.19*	10.76±0.20
S, R→ C	6.74±0.20	6.83±0.20	3.37±0.19	12.86±0.21	10.34±0.20	11.57±0.21
P, R→ C	9.69±0.18	10.40±0.18	2.37±0.17	15.43±0.20	10.52±0.19	12.64±0.22
Q, I→ P	15.35±0.25	17.42±0.25	1.89±0.23	2.76±0.24*	3.42±0.24*	**2.03**±0.23*
Q, S→ P	21.61±0.27	21.60±0.27	3.06±0.25	2.24±0.22*	**2.21**±0.22*	2.60±0.25*
Q, C→ P	19.89±0.26	22.82±0.26	3.01±0.24	3.01±0.25*	2.75±0.25*	**2.50**±0.24*
Q, R→ P	12.37±0.28	16.34±0.28	2.17±0.26	6.23±0.28*	**4.70**±0.27*	5.73±0.28*
I,S→ P	14.93±0.27	12.40±0.27	2.88±0.25	3.22±0.26*	2.67±0.26*	**2.40**±0.24*
I, C→ P	22.73±0.28	19.30±0.28	1.95±0.24	**1.88**±0.24*	6.02±0.27*	9.58±0.28*
I, R→ P	12.76±0.28	13.02±0.28	3.36±0.27	13.46±0.29	**7.67**±0.28*	18.22±0.32
S, C→ P	19.35±0.28	19.09±0.28	3.39±0.27	**2.34**±0.25*	4.77±0.28*	3.40±0.26*
S, R→ P	11.44±0.28	11.53±0.28	1.57±0.23	11.06±0.28	**2.12**±0.25*	9.35±0.30*
C, R→ P	13.89±0.28	14.76±0.28	3.49±0.27	12.50±0.28*	**5.43**±0.28*	15.03±0.32
Q, I→ R	10.47±0.36	12.56±0.36	2.17±0.31	7.49±0.37*	8.29±0.37*	**4.22**±0.35*
Q, S→ R	16.68±0.36	16.68±0.36	2.51±0.33	8.07±0.35*	**6.68**±0.35*	7.12±0.36*
Q, C→ R	13.41±0.37	16.17±0.37	2.26±0.33	8.97±0.36*	10.75±0.37*	**6.19**±0.36*
Q, P→ R	10.78±0.35	12.41±0.35	1.98±0.30	10.66±0.36*	10.07±0.38*	**5.13**±0.35*
I, S→ R	11.37±0.38	9.06±0.38	2.65±0.33	11.36±0.39	**6.32**±0.36*	10.92±0.39
I, C→ R	13.12±0.38	10.11±0.38	2.52±0.33	16.17±0.39	12.36±0.37	**7.63**±0.37*
I, P→ R	10.11±0.35	7.32±0.35	2.11±0.30	19.77±0.37	9.15±0.36	24.13±0.41
S, C→ R	12.52±0.38	12.28±0.38	2.30±0.31	15.17±0.38	12.65±0.37	**8.84**±0.37*
S, P→ R	10.10±0.35	7.72±0.35	2.41±0.29	18.23±0.37	4.85±0.34*	9.53±0.38
C, P→ R	9.59±0.35	8.10±0.35	2.38±0.31	17.94±0.37	12.34±0.37	11.76±0.39

Table S.5: Quantitative evaluation of calibration on the DomainNet dataset [3] for experiments using Clipart (C), Painting (P) or Real (R) as the target domain. For each experiment, we evaluate each of the three proposed algorithms over 1000 experiments with randomly selected test data from the target domain. Results with statistically significant improvement against source-only method are highlighted with asterisks.

		A→ C	P→ C	R→ C	C→ A	P→ A	R→ A	C→ P	A→ P	R→ P	C→ R	A→ R	P→ R
Uncalibrated	Avg.	11.84	15.81	16.58	7.61	12.52	7.80	5.78	6.81	4.38	5.86	4.31	4.59
	2.5%	10.26	14.2	14.92	6.66	11.5	6.81	4.31	5.28	3.16	4.55	3.08	3.38
	97.5%	13.36	17.42	18.23	8.64	13.5	8.72	7.33	8.36	5.64	7.17	5.55	5.86
Source-only	Avg.	16.95	20.3	16.82	7.37	17.05	7.96	5.57	10.64	4.50	5.75	5.34	7.18
	2.5%	15.39	18.7	15.17	6.41	16.00	7.00	4.08	8.94	3.28	4.48	4.03	5.84
	97.5%	18.45	21.94	18.46	8.41	18.02	8.90	7.09	12.18	5.79	7.07	6.68	8.57
Target-only (oracle)	Avg.	4.3	4.24	3.76	4.08	3.43	3.16	3.32	5.35	3.63	3.67	4.19	4.18
	2.5%	2.98	2.99	2.51	3.13	2.56	2.39	2.25	3.89	2.51	2.51	2.97	2.86
	97.5%	5.64	5.71	5.10	5.06	4.39	3.97	4.55	6.74	4.82	4.97	5.42	5.57
TransCal [6]	-	22.9	40.4	4.5	21.7	18.5	21.6	14	9.3	15.6	6.4	5.1	13.9
WTS [2]	-	12.8	26.8	17.3	6.9	8.5	10.4	6.4	1.5	3.8	5.7	6.4	10.8
Set-level	Avg.	10.98	7.71	11.19	4.50	3.68	4.44	3.22	5.39	8.1	3.73	7.86	10.83
	2.5%	9.39	6.24	9.56	3.68	2.82	3.53	2.15	3.99	6.74	2.57	6.37	9.27
	97.5%	12.49*	9.29*	12.81	5.48*	4.64*	5.41*	4.45	7.03	9.65	5.06	9.49	12.29
Cluster-level: NN	Avg.	12.54	9.12	12.64	4.72	6.92	3.05	3.50	6.26	5.54	4.00	6.75	8.72
	2.5%	10.96	7.56	11.06	3.81	5.91	2.22	2.44	4.82	4.23	2.83	5.35	7.20
	97.5%	14	10.94*	14.21	5.69*	7.96*	3.93*	4.73	7.89	6.99	5.32	8.12	10.15
Cluster-level: Regression	Avg.	13.1	10.43	12.48	4.48	5.46	3.90	3.95	6.08	5.30	3.95	6.08	8.05
	2.5%	11.55	8.85	10.84	**3.54**	4.49	3.15	2.74	4.55	4.08	2.72	4.65	6.61
	97.5%	14.63	12.19*	14.06	5.38*	6.42*	4.78*	5.19	7.54	6.58	5.28	7.54	9.64
Ensemble (Avg. logits)	Avg.	12.53	9.10	12.28	5.02	4.36	3.04	3.25	5.57	5.27	3.82	6.25	7.88
	2.5%	11.01	7.56	10.69	4.15	3.44	**2.32**	2.18	4.16	3.90	2.60	4.79	6.37
	97.5%	14.01	10.70	13.94	5.99	5.26	3.85*	4.37	7.05	6.69	5.17	7.77	9.32

Table S.6: Confidence intervals of ECE (%) on Office-Home [4]. Results with statistically significant improvement against source-only and domain adaptation methods are highlighted with asterisks.

		A→ C	P→ C	R→ C	C→ A	P→ A	R→ A	C→ P	A→ P	R→ P	C→ R	A→ R	P→ R
Source-only	Avg.	-1.88	-2.07	-0.10	0.05	-1.77	-0.05	0.04	-0.86	-0.01	0.04	-0.14	-0.48
	2.5%	-2.08	-2.26	-0.12	0.05	-1.90	-0.05	0.02	-1.07	-0.02	0.02	-0.35	-0.65
	97.5%	-1.67	-1.88	-0.09	0.06	-1.64	-0.04	0.05	-0.63	-0.01	0.05	-0.07	-0.33
Target-only (oracle)	Avg.	1.51	3.22	3.37	0.70	2.19	0.75	0.31	0.24	0.09	0.34	0.01	-0.04
	2.5%	0.93	2.57	2.62	0.46	1.84	0.55	0.03	0.06	0.00	0.12	-0.00	-0.27
	97.5%	2.04	3.93	4.12	0.94	2.51	0.93	0.57	0.40	0.18	0.57	0.02	0.21
Set-level	Avg.	0.25	2.55	1.89	0.61	2.19	0.51	0.32	0.12	-0.97	0.30	-0.90	-1.53
	2.5%	0.21*	2.22*	1.64*	0.49*	1.83*	0.19*	0.07*	-0.19*	-1.44	0.02	-1.23	-2.03
	97.5%	0.28	2.91	2.13	0.73	2.52	0.80	0.57	0.42	-0.50	0.59	-0.59	-0.96
Cluster-level: NN	Avg.	-0.21	2.41	1.55	0.58	1.03	0.68	0.59	0.08	-0.77	0.46	-0.63	-1.21
	2.5%	-0.52*	2.14*	1.32*	0.46*	0.78*	0.44*	0.31*	-0.24*	-1.19	0.22*	-0.96	-1.67
	97.5%	0.09	2.73	1.78	0.72	1.25	0.90	0.86	0.41	-0.37	0.72	-0.33	-0.73
Cluster-level: Regression	Avg.	-0.39	2.30	2.15	0.82	1.16	0.84	0.80	0.05	-0.65	0.38	-0.70	-1.70
	2.5%	-0.82*	1.96*	1.80*	0.63*	0.86*	0.57*	0.50*	-0.40*	-1.07	0.06*	-1.06	-2.26
	97.5%	0.03	2.64	2.52	1.01	1.45	1.11	1.12	0.50	-0.25	0.72	-0.34	-1.14
Ensemble: (avg. logits)	Avg.	0.05	2.52	1.94	0.76	1.74	1.02	0.72	0.34	-0.45	0.47	-0.47	-1.04
	2.5%	-0.18*	2.24*	1.71*	0.63*	1.48*	0.79*	0.47*	0.02*	-0.84	0.22*	-0.74	-1.50
	97.5%	0.27	2.83	2.19	0.89	1.99	1.24	0.97	0.66	-0.08	0.75	-0.20	-0.58

Table S.7: Confidence intervals of calibration gain [7] (%) on Office-Home. Results with statistically significant improvement (higher gain) against source-only method are highlighted with asterisks.

References

[1] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 1

[2] Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. In *arXiv:2006.16405*, 2020. 4

[3] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Int. Conf. Comput. Vis.*, 2019. 2, 3

[4] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 4

[5] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. Starcraft II: A new challenge for reinforcement learning. In *arXiv, 1708.04782*, 2017. 1

[6] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Transferable calibration with lower bias and variance in domain adaptation. In *NeurIPS*, 2020. 1, 4

[7] Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, 2020. 4