Supplementary material Generality-training of a Classifier for Improved Calibration in Unseen Contexts

Bhawani Shankar Leelar[0009−0006−2580−1498] and
Meelis Kull[0000−0001−9257−595X]

Institute of Computer Science, University of Tartu, Tartu, Estonia
{bhawani.shankar.leelar,meelis.kull}@ut.ee

A CIFAR-10-C dataset

Here we give detailed description of CIFAR-10-C dataset.

CIFAR-10-C We used CIFAR-10 dataset [1] and apply 15 corruptions [2, 3] from level 1 to level 5 on it, where levels show the severity of a corruption with level 1 being less sever and level 5 means more sever corruption. We consider 4 corruptions (Gaussian Noise, Brightness and Pixelate, Gaussian blur) and original images as source domains, 4 corruptions (Fog, Contrast, Elastic Transform and Saturate) as calibration domains and 7 corruptions (Shot Noise, Impulse Noise, Defocus Blur, Glass Blur, Zoom Blur, JPEG Compression and Speckle Noise) as target domains. The CIFAR-10 dataset contains 60000 images (6000 images per class) of size 32X32, split into 50000 training images and 10000 test images. We randomly sample 200 images without replacement from training part and apply 8 corruption of different levels to get training and validation sets. Rest of the images are used as original for training along with 4 corruptions. This splits make sure that an image appears only once in either original or any corruptions of train or valid set to get a sense of distribution shift. The train part has 26800 original images and 200 images each of 4 different corruption at all 5 levels. The calibration set contains 200 images of 8 corruptions of each level and 200 original images. The test set contains 10000 images from original test split along with 15 versions of corruptions applied on each image from level 1 to 5. In the main paper we have shown results for calibration performance while considering source domains into calibration domains during calibration, here will also show results when only calibration domains are considered during calibration.

B Resources and model parameters

B.1 Computing resource

We used NVidia Graphics Card Tesla V100 and 32 GB of RAM on High Performance Computing (HPC) Cluster for model training, calibration and evaluating the performance on CIFAR-10-C and Office-Home datasets while Tesla A100

with 80GB RAM on DomainNet dataset. We used ResNet101 pre-trained on ILSVRC-1000 and retrained for each dataset.

B.2 Hyper-parameters

CIFAR-10-C We found best value of ρ equal to 0.4 by 3-fold cross validation for both settings: (i) When source domains are considered for calibration along with calibration domains and (ii) when only calibration domains are considered for calibration.

Office-Home Best values of ρ for Office-Home dataset are given in Table 1. These values are selected by 3-fold cross validation for best error.

Table 1: Value of best ρ when we consider source domain into calibration domains (Source±Calib.) and when we consider only calibration domains (Calib.) for Office-Home dataset

Source Domains	Calibration Domains	ρ (ResNet)		ρ (EfficientNet)	
		Source+Calib.	Calib.	Source+Calib.	Calib.
Art	{Product, Real World}	0.2	0.2	0.2	0.2
Art	{Clipart, Real World}	0.2	0.2	0.2	0.2
Art	{Clipart, Product}	0.2	0.2	0.2	0.2
Clipart	{Product, Real World}	0.3	0.3	0.2	0.2
Clipart	{Art, Real World}	0.3	0.3	0.2	0.3
Clipart	{Art, Product}	0.3	0.3	0.2	0.2
Product	{Clipart, Real World}	0.2	0.2	0.2	0.2
Product	{Art, Real World}	0.2	0.2	0.4	0.2
Product	{Art, Clipart}	0.3	0.3	0.2	0.3
Real World	{Clipart, Product}	0.2	0.2	0.2	0.2
Real World	{Art, Product}	0.3	0.3	0.3	0.2
Real World	{Art, Clipart}	0.2	0.2	0.3	0.2

DomainNet We found the best value of ρ at 0.2 for all the combination of source and calibration domains except for three combinations where we found it equal to 0.3. In first setting when we consider source domains in calibration domains, there are two combinations: (i) source domains: {Clipart, Real}, calibration domains: {Infograph, Painting, Sketch} and (ii) source domains: {Quickdraw, Sketch}, calibration domains: {Clipart, Infograph, Painting}, where we used $\rho = 0.3$. In second setting when we consider only calibration domains for calibration and found one combination: source domains: {Quickdraw, Real}, calibration domains: {Clipart, Infograph, Painting}, where we used $\rho = 0.3$. We used 3-fold cross validation based on best error to select best ρ.

C Focal loss experiments

We also tested the method using focal loss [4] for training. We also used focal loss for generality-training replacing negative log likelihood (NLL) (convex sum of KL divergence and focal loss) in the same manner used for main experiments.

C.1 Calibration performance

Calibration performance is similar as obtained by NLL in main paper. The results shown in Tables 2 and 3 are consistent with results shown in Tables 7 and 10 of the main results. Focal loss [4] known to generating better calibrated results but it does not provide calibration in unseen domains. We observe that both cluster based methods [5] and CaliGen improve the results when calibration is learned on model trained with focal loss.

Table 2: Calibration performance (ECE %) evaluated on target domains of Office-Home dataset and averaged by target domains while model trained using focal loss and CaliGen trained using focal loss and KL divergence

Method	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated		38.97±3.66	40.73±4.11	29.52±2.17	25.54±5.6	33.69±7.55
TS	Source Only	17.67±6.11	23.88±4.94	11.39±6.53	10.9±1.61	15.96±7.4
TS	*Oracle*	5.25±1.39	6.83±1.5	7.3±0.64	7.47±1.96	6.71±1.7
TS	Source and Calibration Domains	8.73±3.18	14.81±1.61	9.02±1.91	9.17±2.36	10.43±3.45
CPCS		9.79±4.56	15.05±2.96	8.14±1.57	9.26±3.11	10.56±4.18
TransCal		18.66±19.9	30.32±15.1	13.16±9.69	27.87±4.66	22.5±15.26
HB-TL		26.61±1.69	30.58±3.32	20.65±1.92	15.42±2.65	23.31±6.28
Cluster NN		8.06±1.69	15.04±3.08	8.63±1.98	8.25±1.69	9.99±3.65
Cluster LR		8.13±0.72	13.93±4.48	8.73±1.71	8.57±1.4	9.84±3.46
Cluster En.		8.08±1.88	14.78±3.01	8.38±1.98	8.24±1.43	9.87±3.56
CaliGen		**6.91±1.94**	**12.1±2.99**	**7.85±2.19**	**6.12±0.42**	**8.24±3.12**
CaliGen TS		28.69±4.67	35.86±4.14	22.59±4.91	20.8±0.47	26.98±7.12
CaliGen En.		6.97±0.48	12.68±2.43	9.48±3.27	9.08±1.16	9.55±2.95
TS	Calibration Domains Only	7.58±2.51	11.98±0.65	11.97±5.17	12.81±4.93	11.08±4.32
CPCS		8.42±4.09	11.65±2.44	11.16±5.34	13.93±6.39	11.29±5.18
TransCal		8.52±4.05	26.85±13.12	10.62±4.99	8.36±1.84	13.59±10.66
HB-TL		26.59±5.32	32.57±7.08	24.04±1.39	17.78±2.39	25.24±7.05
Cluster NN		6.54±0.74	12.65±1.53	10.55±5.13	11.35±3.9	10.27±4.04
Cluster LR		6.14±1.11	12.36±2.76	11.31±4.04	10.37±2.62	10.05±3.69
Cluster En.		6.63±0.84	12.39±1.68	10.44±5.17	10.81±3.99	10.07±4.0
CaliGen		**5.52±0.77**	10.4±1.5	**9.15±2.92**	**6.67±0.61**	**7.94±2.59**
CaliGen TS		31.54±1.25	37.23±0.66	22.73±4.85	19.43±2.75	27.73±7.61
CaliGen En.		5.89±0.94	**9.59±1.05**	13.7±4.83	14.61±3.2	10.94±4.58

Table 3: Calibration performance (ECE %) evaluated on source domains of Office-Home dataset and averaged by source domains with same settings as Table 2

Method	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated		0.19±0.0	0.47±0.0	0.1±0.0	0.59±0.0	0.34±0.2
TS	Source Only	0.23±0.0	2.33±0.0	0.18±0.0	7.18±0.0	2.48±2.85
TS	*Oracle*	0.19±0.0	0.49±0.0	0.1±0.0	0.38±0.0	0.29±0.15
TS	Source and Calibration Domains	**1.56±0.75**	5.74±1.11	2.2±1.18	17.18±5.01	6.67±6.81
CPCS		2.02±1.12	4.18±0.87	1.97±0.99	14.45±4.31	5.66±5.65
TransCal		19.62±31.52	**0.48±0.01**	**0.69±1.02**	14.08±3.79	8.72±17.95
HB-TL		12.98±2.21	13.29±3.1	14.13±3.07	14.28±3.77	13.67±3.13
Cluster NN		2.19±1.12	4.37±0.29	2.24±1.19	15.02±3.67	5.96±5.68
Cluster LR		3.06±1.58	3.94±0.38	2.49±1.19	13.48±2.64	5.74±4.79
Cluster En.		1.62±0.77	4.16±0.22	2.02±1.08	14.45±3.5	5.56±5.55
CaliGen		5.22±0.39	11.07±0.36	4.57±0.23	28.12±3.83	12.24±9.7
CaliGen TS		3.77±0.56	2.39±0.32	0.91±0.16	**6.21±1.02**	**3.32±2.04**
CaliGen En.		3.53±1.28	7.55±0.84	2.3±0.62	24.64±5.04	9.51±9.34
TS	Calibration Domains Only	2.69±1.54	8.76±2.71	5.32±4.06	24.66±9.46	10.36±10.09
CPCS		4.27±2.71	6.63±2.29	6.2±5.06	22.34±10.3	9.86±9.42
TransCal		**0.47±0.32**	**3.53±0.99**	**1.7±1.83**	14.33±4.83	**5.01±6.09**
HB-TL		9.95±2.3	9.27±1.51	9.12±3.34	6.42±3.32	8.69±3.04
Cluster NN		2.82±1.59	6.06±1.22	5.07±3.72	21.29±7.0	8.81±8.37
Cluster LR		3.5±1.69	4.52±1.39	4.88±3.37	19.65±5.1	8.14±7.41
Cluster En.		2.24±1.23	5.1±0.97	4.41±3.22	21.17±6.98	8.23±8.5
CaliGen		10.77±1.88	18.06±6.01	13.59±2.04	25.98±6.96	17.1±7.49
CaliGen TS		4.73±2.87	7.22±3.08	3.48±2.37	**5.49±1.07**	5.23±2.82
CaliGen En.		7.56±4.26	19.59±4.7	10.97±7.33	33.44±9.47	17.89±12.07

C.2 Error generalization

With focal loss also we achieved improved error generalization on target domains as shown in Table 4.

D Additional results

D.1 Calibration performance on target domains

Here we present results evaluated on target domains for each dataset for two settings: (i) we use source domains along with calibration domains for calibration and (ii) we only use calibration domains for calibration. We have shown all the results related with first settings in main paper, but here we present again for better readability along with second setting results. We also show results for CaliGen Ensemble which is calculated by taking mean of logits obtained by

Table 4: Error % on Office-Home averaged by target domain where calibration domains (in) include source domain, (out) does not include source domain with same settings as Table 2

Method	Art	Clipart	Product	RealWorld	Average
Uncalibrated	78.05±4.07	75.46±3.25	63.43±8.5	59.44±6.3	69.1±9.81
CaliGen (in)	**74.72±2.01**	**73.05±2.24**	60.25±1.37	**53.72±5.05**	65.43±9.28
CaliGen Ensem (in)	74.77±2.38	73.13±2.23	**59.11±5.11**	53.81±5.51	**65.2±9.85**
CaliGen (out)	75.93±2.57	74.53±1.98	62.76±1.11	57.82±4.36	67.76±8.17
aliGen Ensem (out)	74.89±2.1	73.25±2.44	59.94±4.19	53.62±5.14	65.42±9.68

CaliGen and Temperature Scaling (TS) method (first method in each setting). We also show results for CaliGen (TS) where TS is applied on CaliGen logits of instances from validation set (calibration domain).

Results from Tables 5 to 7 and 9 clearly indicate that we should use source domains along with calibration domains for calibration to achieve better calibrated probabilities. Applying TS on top of CaliGen might overfit the model to the source and calibration domains as evident from calibration performance on source domains given in next sub-section.

D.2 Calibration performance on source domains

Results for source domain Expected Calibration Error (ECE) are shown in Tables 10 and 12 and for datasets Office-Home and DomainNet, respectively. We observe that TS on top of CaliGen gives best results on both cases. The temperature obtained on CaliGen logits is less than 1 (in the range of 0.65 to 0.95 in most cases) for $0.2 \leq \rho \leq 0.8$, which indicates that CaliGen makes the model under-confident while fine-tuning with calibration domains. Applying TS on top of CaliGen makes the model better calibrated for source domain by increasing the confidence. However, on target domains, the model is better calibrated without applying TS as model has learned to produce lower confidence which helps in achieving better calibrated probabilities. Source domain results for CIFAR-10-C can be inferred from the column 'Train Filter' (source domains) of Tables 5 and 6.

D.3 Cross-entropy loss on target domains

We show cross-entropy loss evaluated on target domains for each datasets in Tables 13 to 15 and 17. It is obvious that CaliGen achives better loss compared to cluster based methods as we improve on error. On all datasets, CaliGen achives better loss compared to Cluster based methods however, in some cases CaliGen ensemble beats CaliGen.

Table 5: Calibration performance (ECE %) for model trained on original images and 4 filters with corruption level 1, where column Train Filter and Calib. Filter represents the source and calibration domains while Test Filter (1) represents the rest of the test filter of severity of corruption 1 and All Filter (i) represents all the filters with severity of corruption i

Method	Train Filter	Calib. Filter	Test Filter (1)	All Filter (2)
Uncalibrated	14.32±2.96	15.16±2.64	16.65±3.31	20.48±5.43
TS Source	5.77±1.35	6.02±1.43	6.56±1.1	9.12±3.81
TS Oracle	5.22±0.33	5.3±0.31	5.6±0.2	5.56±0.58
Calibrated on Source and Calibration Domains				
HB-TL	7.7±1.54	7.82±1.75	7.93±2.81	10.02±2.62
TS	5.62±0.47	5.72±0.47	5.72±0.42	7.87±3.04
CPCS	5.73±0.21	5.76±0.22	5.65±0.29	7.38±2.52
TransCal	5.7±0.17	5.73±0.17	5.68±0.31	7.43±2.61
Cluster NN	5.63±1.12	5.88±1.12	6.0±0.85	8.56±3.72
Cluster LR	16.57±1.01	16.49±1.15	17.6±2.57	20.25±4.34
Cluster En.	9.22±1.39	9.22±1.61	10.21±2.13	12.84±4.06
CaliGen	**5.37±0.82**	**5.5±0.87**	5.69±0.46	**7.36±3.18**
CaliGen TS	7.88±1.9	8.26±1.96	8.82±1.24	11.61±4.25
CaliGen En.	5.43±0.65	5.54±0.7	**5.42±0.4**	7.52±3.1
Calibrated only on Calibration Domains				
HB-TL	11.48±1.97	11.72±2.19	11.72±3.18	15.9±5.77
TS	**5.99±0.28**	**6.01±0.32**	**5.68±0.43**	**7.03±2.08**
CPCS	6.33±0.54	6.32±0.62	5.73±0.74	6.82±1.93
TransCal	8.69±1.62	8.47±1.79	7.56±1.46	7.05±2.01
Cluster NN	6.01±1.16	6.15±1.29	6.26±0.72	8.81±4.13
Cluster LR	12.64±0.58	12.51±0.56	11.94±0.61	13.93±2.56
Cluster En.	7.58±0.84	7.57±0.97	7.36±0.52	9.91±3.37
CaliGen	8.95±2.86	9.49±2.99	10.11±1.91	13.68±5.4
CaliGen TS	17.51±3.8	18.5±3.57	19.65±3.17	23.95±6.0
CaliGen En.	5.93±1.77	6.2±1.92	6.46±1.11	9.05±4.34

D.4 Error performance on target domains

Error results shown in Tables 18 to 20 and 22 for datasets CIFAR-10-C, Office-Home and DomainNet, respectively. We achieve better error on all three datasets, however here again we observe that it is better to calibrate on both source and calibration domains for better performance.

D.5 Temperature parameter for TS based methods

We show the value of temperature in Tables 23 to 26 for each TS based method and show that on CIFAR-10-C dataset, the temperature attained by Cluster LR is negative which makes the higher error for this method. The ideal temperature should be close to Oracle.

Table 6: Calibration performance (ECE %) for model trained on original images and 4 filters with corruption level 1, where column Train Filter and Calib. Filter represents the source and calibration domains while Test Filter (1) represents the rest of the test filter of severity of corruption 1 and All Filter (i) represents all the filters with severity of corruption i

Method	All Filter (3)	All Filter (4)	All Filter (5)	Average
Uncalibrated	27.03±9.97	32.07±13.61	35.73±14.86	23.06±11.99
TS Source	14.57±8.38	18.86±12.34	21.69±14.08	11.8±10.0
TS Oracle	5.69±0.95	5.71±1.05	5.23±1.2	5.47±0.78
Calibrated on Source and Calibration Domains				
HB-TL	12.91±3.48	14.98±4.08	**16.44±3.57**	11.11±4.51
TS	12.97±7.39	16.88±11.55	19.54±13.33	10.62±9.11
CPCS	12.1±6.85	15.89±11.07	18.45±12.94	10.14±8.58
TransCal	12.23±6.96	16.05±11.15	18.63±13.01	10.21±8.67
Cluster NN	13.86±8.5	17.95±12.93	20.66±14.87	11.22±10.09
Cluster LR	25.61±7.51	29.87±11.17	32.44±13.61	22.69±9.69
Cluster En.	18.15±8.19	22.43±11.97	25.13±13.89	15.31±9.93
CaliGen	**11.6±7.23**	**14.97±11.04**	17.3±12.71	**9.68±8.41**
CaliGen TS	16.53±8.71	20.32±12.41	23.1±13.91	13.79±9.87
CaliGen En.	12.18±7.28	15.84±11.48	18.54±13.24	10.07±8.87
Calibrated only on Calibration Domains				
HB-TL	21.78±10.66	26.24±14.86	29.54±16.29	18.34±11.91
TS	11.33±6.19	15.04±10.41	17.43±12.37	9.79±7.98
CPCS	11.01±5.7	14.61±9.99	16.83±12.06	**9.66±7.62**
TransCal	**9.66±3.92**	**12.86±7.95**	**14.65±10.07**	9.85±5.86
Cluster NN	14.14±9.29	18.29±14.06	20.92±16.2	11.51±10.71
Cluster LR	18.1±6.57	21.95±11.23	24.18±14.23	16.46±8.69
Cluster En.	14.88±7.97	18.89±12.52	21.38±14.83	12.51±9.71
CaliGen	17.99±9.51	20.68±11.74	22.47±12.4	14.77±9.42
CaliGen TS	28.94±10.04	32.25±12.25	34.48±13.01	25.04±10.55
CaliGen En.	13.51±8.65	16.68±11.71	18.54±13.02	10.91±9.08

D.6 Source domains results of model trained on calib domains

We show results obtained by model trained source and calibration domains averaged by target and source domains in Table 27. When compared with results of CaliGen in Table 7 for target domains and in Table 10 for source domain, it is clear that CaliGen gives comparable performance on ECE for target domains while beating this stronger model on source domain calibration performance.

References

[1] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. "CIFAR-10 (Canadian Institute for Advanced Research)". In: (2009). URL: http://www.cs.toronto.edu/~kriz/cifar.html.

Table 7: Calibration performance (ECE %) evaluated on target domains of Office-Home dataset and averaged by target domains

Method	Calibrated on	Art	Clipart	Product	Real World	Average
Uncalibrated		37.61±5.21	40.32±0.28	29.64±4.81	26.05±7.48	33.41±7.75
TS	Source Only	18.12±5.42	24.44±5.38	11.64±5.34	11.25±0.74	16.36±7.15
TS	Oracle	*4.77±0.51*	*6.06±0.88*	*6.6±0.9*	*6.59±1.13*	*6.0±1.16*
HB-TL	Source and Calibration Domains	25.38±1.73	25.99±4.87	18.57±5.92	14.68±3.54	21.16±6.4
TS		8.24±3.15	15.87±1.4	7.73±3.09	8.77±2.91	10.15±4.3
CPCS		8.98±3.08	15.47±1.62	7.26±1.88	9.32±3.6	10.26±4.1
TransCal		28.76±14.78	23.91±15.78	18.2±10.44	15.5±8.38	21.6±13.71
Cluster NN		8.03±1.79	17.08±2.1	8.2±2.9	7.92±1.65	10.31±4.47
Cluster LR		7.4±0.86	17.79±3.33	8.56±2.81	7.48±1.7	10.31±4.95
Cluster En.		7.77±1.81	17.15±2.02	7.84±2.9	7.56±2.07	10.08±4.66
CaliGen		**6.64±1.84**	14.61±4.87	**6.38±0.93**	**7.27±0.54**	**8.72±4.32**
CaliGen TS		29.39±0.7	35.9±1.91	23.2±2.14	18.53±2.26	26.75±6.8
CaliGen En.		6.77±0.64	**13.8±2.87**	7.7±2.28	9.71±1.64	9.49±3.38
HB-TL	Calibration Domains only	28.24±3.53	32.81±8.18	23.33±2.83	20.06±4.7	26.11±7.13
TS		6.32±1.11	12.99±1.23	10.71±1.58	12.89±1.37	10.73±3.01
CPCS		6.39±2.83	11.93±1.45	10.17±5.41	14.65±6.66	10.78±5.47
TransCal		18.86±15.04	24.22±14.03	11.28±7.12	17.8±7.37	18.04±12.38
Cluster NN		6.31±0.99	15.09±1.27	9.49±1.3	10.94±1.42	10.46±3.39
Cluster LR		6.39±1.11	17.45±1.25	9.57±1.28	**9.91±1.3**	10.83±4.25
Cluster En.		6.25±1.09	15.49±1.21	9.59±1.41	10.62±1.36	10.49±3.55
CaliGen		6.81±1.2	11.48±1.55	**6.56±1.3**	10.14±1.3	**8.75±2.51**
CaliGen TS		28.41±1.72	36.43±1.71	21.89±1.74	17.27±1.36	26.0±7.39
CaliGen En.		**5.54±0.97**	**9.9±1.31**	11.15±1.69	15.8±1.42	10.6±3.9

[2] Dan Hendrycks and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=HJz6tiCqYm.

[3] TensorFlow Datasets Team. *imagenet2012_corrupted*. 2020 [Online]. URL: https://github.com/tensorflow/datasets/blob/master/tensorflow_datasets/image_classification/imagenet2012_corrupted.py.

[4] Jishnu Mukhoti et al. "Calibrating Deep Neural Networks using Focal Loss". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 15288–15299. URL: https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf.

[5] Yunye Gong et al. "Confidence Calibration for Domain Generalization under Covariate Shift". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 8958–8967.

Table 8: Calibration performance (ECE %) evaluated on target domains of Office-Home dataset and averaged by target domains when the classifier trained using EfficientNet V2 B0

Method	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated		40.83±5.8	35.66±8.22	28.77±4.66	27.75±7.27	33.25±8.5
TS	Source Only	16.49±8.3	17.91±7.52	9.95±2.49	10.58±2.99	13.73±6.89
TS	Oracle	5.46±0.77	5.81±1.02	6.29±0.69	5.71±0.13	5.82±0.79
HB-TL	Source and Calibration Domains	26.31±8.99	25.91±12.5	15.7±5.61	17.54±5.37	21.36±9.86
TS		9.02±2.39	10.82±2.38	7.07±0.97	5.75±0.24	8.17±2.61
CPCS		8.11±1.77	9.43±1.85	7.02±1.31	5.79±0.47	7.59±1.98
TransCal		29.57±19.09	23.15±18.02	15.12±12.23	21.19±7.47	22.26±15.82
Cluster NN		9.24±2.71	11.59±3.1	7.06±0.86	5.71±0.38	8.4±3.07
Cluster LR		9.28±2.3	12.01±3.36	**7.0±0.85**	5.87±0.49	8.54±3.15
Cluster En.		9.13±2.54	11.56±2.98	7.12±0.9	**5.52±0.64**	8.33±3.04
CaliGen		**6.06±1.29**	**6.78±1.94**	7.77±3.33	6.59±1.61	**6.8±2.27**
CaliGen TS		25.38±3.93	25.76±4.1	17.76±4.48	17.74±1.67	21.66±5.4
CaliGen En.		6.89±1.45	8.43±2.59	7.34±2.61	6.99±1.29	7.41±2.17
HB-TL	Calibration Domains only	31.83±10.85	29.84±12.98	20.17±6.45	23.26±7.52	26.27±10.89
TS		7.2±1.32	8.58±0.54	8.11±3.29	8.37±4.03	8.07±2.75
CPCS		6.42±0.54	7.02±0.8	8.79±4.24	9.03±4.47	**7.82±3.32**
TransCal		20.46±14.86	19.38±19.93	**6.6±0.49**	17.05±5.64	15.87±13.88
Cluster NN		7.54±1.27	9.58±1.11	7.95±3.24	8.02±3.38	8.27±2.61
Cluster LR		7.29±1.47	10.44±2.16	8.01±3.02	**7.62±2.43**	8.34±2.64
Cluster En.		7.38±1.45	9.62±1.25	7.92±3.09	7.92±3.28	8.21±2.59
CaliGen		**4.88±0.57**	**5.95±0.32**	9.45±2.78	12.16±4.16	8.11±3.83
CaliGen TS		25.3±1.03	28.74±3.35	19.06±1.39	16.34±2.82	22.36±5.45
CaliGen En.		5.77±0.43	6.32±0.65	10.25±5.12	12.58±4.87	8.73±4.53

Table 9: Calibration performance (ECE %) evaluated on target domains of DomainNet dataset and averaged by target domains

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
Uncalibrated	16.12±3.2	28.95±6.9	22.74±5.24	38.33±13.88	18.61±3.71	22.1±3.73	24.48±10.25
TS (Source Only)	10.98±2.83	22.31±5.58	15.39±3.99	28.43±13.09	11.26±4.01	14.05±2.05	17.07±9.0
Oracle (TS)	5.93±0.68	4.72±0.73	4.96±0.78	2.44±0.69	5.08±0.6	5.34±0.37	4.74±1.28
Calibrated on Source and Calibration Domains							
HB-TL	6.99±1.03	17.3±2.92	10.83±2.29	22.34±3.2	7.66±1.72	10.16±2.02	12.55±5.97
TS	8.02±2.54	10.35±4.34	6.07±1.48	18.23±11.84	9.06±3.66	6.11±1.25	9.64±6.89
CPCS	7.09±1.54	13.02±5.35	7.44±3.07	21.5±12.52	8.02±1.33	6.35±1.82	10.57±7.9
TransCal	6.94±1.61	19.39±7.9	11.05±3.73	23.7±13.04	7.2±2.25	9.07±3.23	12.89±9.21
Cluster NN	7.27±1.33	11.67±5.03	6.0±1.76	19.6±8.33	**6.24±1.13**	6.0±1.04	9.46±6.44
Cluster LR	8.43±1.84	13.01±7.69	6.97±2.43	21.86±7.04	7.62±2.16	6.08±0.87	10.66±7.11
Cluster En.	7.2±1.48	11.55±5.34	**5.66±1.18**	20.27±7.61	6.3±1.61	**5.23±0.57**	9.37±6.6
CaliGen	9.63±1.3	**6.91±1.81**	5.83±0.62	**12.17±1.56**	8.4±0.86	5.81±0.6	**8.13±2.57**
CaliGen TS	**6.43±0.8**	17.63±3.65	9.99±3.18	23.12±2.25	7.17±1.99	9.46±1.76	12.3±6.53
CaliGen En.	11.25±2.73	7.61±2.9	5.76±1.39	13.42±6.1	10.37±3.42	7.18±1.79	9.27±4.32
Calibrated only on Calibration Domains							
HB-TL	6.6±0.78	14.67±1.75	9.35±1.64	16.94±3.08	**6.67±1.12**	6.84±0.74	10.18±4.48
TS	14.06±5.91	6.78±3.02	7.69±3.71	12.83±8.37	14.29±5.54	9.15±4.43	10.8±6.24
CPCS	8.1±3.44	6.26±1.95	**4.66±1.11**	11.04±3.95	7.17±3.58	8.29±3.75	7.59±3.71
TransCal	**6.39±1.98**	8.51±2.22	5.62±2.53	11.88±5.72	6.79±3.72	**5.73±1.98**	**7.49±3.97**
Cluster NN	10.43±3.15	8.26±4.1	5.29±1.4	15.95±6.69	9.51±3.55	6.27±1.94	9.28±5.19
Cluster LR	11.01±3.8	9.06±5.54	6.93±3.16	21.94±7.18	12.7±5.13	7.75±3.1	11.57±7.0
Cluster En.	10.62±3.12	7.82±4.22	5.57±1.8	17.28±5.82	9.81±3.39	6.77±2.28	9.64±5.3
CaliGen	8.99±1.9	**5.98±1.67**	6.14±1.1	10.87±1.99	8.32±1.09	5.96±1.18	7.71±2.4
CaliGen TS	8.7±0.71	19.39±3.68	10.78±2.16	25.21±2.43	8.72±1.33	11.74±1.85	14.09±6.53
CaliGen En.	17.28±4.79	5.14±2.31	10.78±4.42	**9.85±4.46**	17.02±3.95	11.77±3.55	11.97±5.81

Table 10: Calibration performance (ECE %) evaluated on source domains of Office-Home dataset and averaged by source domains

Domain	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated		0.11±0.0	0.63±0.0	0.1±0.0	2.42±0.0	0.82±0.95
TS	Source Only	0.23±0.0	2.74±0.0	0.18±0.0	13.55±0.0	4.17±5.51
TS	*Oracle*	*0.15±0.0*	*0.58±0.0*	*0.08±0.0*	*0.71±0.0*	*0.38±0.27*
HB-TL	Source and Calibration Domains	14.22±1.99	17.06±3.97	15.0±3.61	20.09±2.38	16.59±3.84
TS		2.21±1.21	7.45±0.85	2.11±1.15	23.48±5.16	8.81±9.16
CPCS		3.24±1.72	6.44±0.47	1.84±0.93	20.64±4.23	8.04±7.82
TransCal		**0.14±0.04**	3.47±4.89	5.76±7.26	14.93±4.82	6.08±7.42
Cluster NN		2.21±1.0	6.68±0.6	2.11±1.08	22.47±5.05	8.37±8.76
Cluster LR		1.98±0.71	5.97±0.62	2.18±1.0	21.5±4.96	7.91±8.41
Cluster En.		1.57±0.69	6.27±0.62	1.7±0.89	21.83±5.22	7.84±8.72
CaliGen		6.92±0.46	11.77±0.44	3.41±0.61	23.12±1.17	11.3±7.48
CaliGen TS		2.08±0.4	**1.96±0.15**	**0.87±0.04**	**4.01±0.23**	**2.23±1.16**
CaliGen En.		3.85±1.59	8.8±0.56	1.95±0.78	24.26±3.64	9.71±9.0
HB-TL	Calibration Domains only	8.95±1.35	8.08±2.92	8.28±1.06	8.14±3.45	8.36±2.44
TS		3.7±2.27	11.53±2.42	5.3±4.22	30.11±9.26	12.66±11.78
CPCS		6.51±3.95	10.1±1.73	6.4±5.53	27.74±9.34	12.69±10.58
TransCal		15.79±24.05	**3.04±1.57**	7.44±12.71	18.95±5.77	11.31±15.31
Cluster NN		3.22±1.62	9.33±1.28	4.58±3.19	28.35±8.54	11.37±11.09
Cluster LR		2.6±1.25	7.26±0.72	4.28±2.38	27.68±8.38	10.45±11.01
Cluster En.		**2.47±1.38**	8.34±1.02	3.69±2.62	28.21±8.89	10.68±11.38
CaliGen		11.3±2.84	14.28±3.89	14.73±7.29	21.32±2.56	15.41±5.84
CaliGen TS		3.71±0.53	7.83±0.94	**3.07±0.69**	**5.61±0.99**	**5.06±2.02**
CaliGen En.		8.68±5.12	20.99±3.95	10.86±7.77	33.04±7.17	18.39±11.47

Table 11: Calibration performance (ECE %) evaluated on source domains of Office-Home dataset and averaged by source domains when classifier is trained using EfficientNet V2 B0

Method	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated		3.77±0.0	0.67±0.0	0.08±0.0	5.1±0.0	2.41±2.09
TS	Source Only	11.16±0.0	3.95±0.0	0.72±0.0	14.95±0.0	7.69±5.64
TS	Oracle	1.65±0.0	0.67±0.0	0.09±0.0	4.38±0.0	1.7±1.65
HB-TL	Source and Calibration Domains	16.46±0.6	14.95±3.18	12.38±2.64	16.97±2.62	15.19±3.04
TS		18.66±2.14	7.02±0.76	5.56±1.56	21.5±2.59	13.18±7.24
CPCS		19.91±1.5	7.78±0.44	6.48±1.09	22.33±2.76	14.12±7.26
TransCal		27.59±14.68	**0.77±0.03**	**0.14±0.09**	18.27±1.42	11.69±13.84
Cluster NN		16.61±1.98	6.88±0.64	5.07±1.29	21.31±2.57	12.47±6.96
Cluster LR		15.93±1.83	5.98±0.31	4.37±1.08	21.05±2.54	11.83±7.12
Cluster En.		16.66±2.0	6.47±0.45	4.78±1.28	21.19±2.59	12.27±7.09
CaliGen		20.63±0.71	11.24±0.15	5.79±0.17	27.22±2.61	16.22±8.39
CaliGen TS		**3.48±0.34**	1.91±0.12	0.65±0.06	**6.24±0.76**	**3.07±2.13**
CaliGen En.		20.83±1.94	7.05±0.44	4.07±0.48	25.47±2.92	14.35±9.18
HB-TL	Calibration Domains only	10.88±0.61	5.67±1.17	6.21±1.08	11.9±3.73	8.67±3.43
TS		20.97±2.91	9.28±1.79	10.53±4.37	25.69±4.83	16.62±7.85
CPCS		22.19±2.26	10.5±1.31	12.69±3.72	27.28±5.81	18.17±7.78
TransCal		33.05±12.64	**4.81±2.8**	**0.14±0.05**	20.02±2.6	14.51±14.57
Cluster NN		18.64±3.23	9.07±1.48	9.33±3.86	25.29±4.48	15.58±7.63
Cluster LR		18.61±4.42	7.78±1.32	7.79±3.21	25.02±4.02	14.8±8.14
Cluster En.		18.95±3.59	8.4±1.29	8.9±3.77	25.17±4.39	15.36±7.86
CaliGen		24.58±2.67	29.65±2.11	26.25±10.39	24.66±2.49	26.28±5.97
CaliGen TS		**3.88±0.23**	5.63±0.7	4.18±1.78	**4.54±0.3**	**4.56±1.18**
CaliGen En.		25.9±3.7	17.37±2.92	16.45±7.83	29.61±5.23	22.33±7.68

Table 12: Calibration performance (ECE %) evaluated on source domains of DomainNet dataset and averaged by source domains

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
Uncalibrated	6.27±0.48	9.0±2.24	7.34±1.33	5.81±0.68	5.57±1.1	8.24±0.48	7.04±1.76
Source Only (TS)	13.88±2.08	17.31±5.42	13.25±1.73	8.22±1.24	9.03±1.56	13.63±3.45	12.55±4.28
Oracle (TS	*4.38±0.61*	*5.29±0.08*	*5.21±0.18*	*4.65±0.48*	*4.17±0.54*	*4.85±0.35*	*4.76±0.58*
Calibrated on Source and Calibration Domains							
HB-TL	24.36±2.51	15.21±4.47	17.69±3.14	14.83±3.58	16.92±4.01	18.38±4.59	17.9±4.93
TS	31.82±5.4	29.22±7.68	30.57±6.41	20.9±4.99	23.44±6.92	29.42±5.73	27.56±7.41
CPCS	26.44±4.86	26.55±7.12	26.52±6.45	15.7±3.51	19.52±7.19	23.58±4.76	23.05±7.13
TransCal	18.75±6.38	22.15±5.22	22.39±6.37	13.29±4.73	14.43±6.94	18.8±5.33	18.3±6.83
Cluster NN	26.49±4.04	22.24±4.03	20.09±3.07	8.69±1.93	16.09±1.83	24.36±4.58	19.66±6.82
Cluster LR	24.54±3.98	19.23±3.57	18.24±2.9	8.15±1.35	14.86±1.78	21.18±3.92	17.7±6.03
Cluster En.	26.41±3.88	23.29±4.53	21.55±2.42	11.17±2.22	16.93±2.17	24.16±4.31	20.59±6.15
CaliGen	22.14±1.01	15.38±2.89	19.1±1.78	14.53±1.01	20.27±1.25	19.32±1.98	18.46±3.21
CaliGen TS	**8.6±1.27**	**6.35±1.35**	**6.91±1.14**	**5.37±0.28**	**8.18±1.24**	**7.25±0.84**	**7.11±1.53**
CaliGen En.	31.47±3.6	26.24±5.23	29.03±3.31	20.46±3.01	24.4±3.21	29.2±4.42	26.8±5.31
Calibrated only on Calibration Domains							
HB-TL	18.61±2.23	5.25±0.48	9.05±2.54	11.0±2.44	8.5±1.95	19.68±3.52	12.01±5.83
TS	40.56±5.73	33.36±8.17	39.27±9.37	34.0±8.1	36.97±9.86	37.91±7.51	37.01±8.64
CPCS	28.76±4.35	13.38±4.34	18.15±4.36	22.88±5.37	18.67±5.37	30.2±5.36	22.01±7.72
TransCal	25.35±3.83	9.92±2.85	14.71±5.65	23.18±7.32	13.12±4.8	28.55±5.97	19.14±8.67
Cluster NN	34.15±4.98	27.68±5.85	27.84±4.87	24.43±3.62	26.53±3.08	30.91±5.52	28.59±5.7
Cluster LR	29.88±5.23	24.95±6.12	24.42±5.36	19.44±5.77	22.99±4.06	26.49±5.1	24.7±6.19
Cluster En.	33.02±4.53	28.5±6.25	28.42±4.22	24.17±4.12	26.7±3.24	29.98±4.72	28.47±5.35
CaliGen	19.28±1.33	7.63±2.13	13.61±1.41	**6.47±1.52**	15.63±2.65	14.16±1.99	12.8±4.85
CaliGen TS	**5.73±0.25**	**8.8±1.82**	**6.59±1.21**	10.53±2.19	**6.18±0.77**	**5.95±0.64**	**7.3±2.21**
CaliGen En.	38.84±4.2	25.02±5.12	33.33±5.48	32.55±4.67	36.82±5.31	33.19±4.77	33.29±6.57

Table 13: Cross-entropy loss for model trained on original images and 4 filters with corruption level 1, where column Train Filter and Calib. Filter represents the source and calibration domains while Test Filter (1) represents the rest of the test filter of severity of corruption 1 and All Filter (i) represents all the filters with severity of corruption i

Method	Train Filter	Calib. Filter	Test Filter (1)	All Filter (2)
Uncalibrated	1.14±0.22	1.2±0.19	1.34±0.31	1.59±0.38
Source Only (TS)	1.01±0.19	1.06±0.17	1.17±0.28	1.37±0.32
Oracle (TS)	*1.0±0.19*	*1.05±0.17*	*1.17±0.27*	*1.34±0.31*
Calibrated on Source and Calibration Domains				
HB-TL	1.31±0.17	1.36±0.16	1.52±0.28	1.83±0.42
TS	1.01±0.19	1.06±0.17	1.17±0.27	1.36±0.31
CPCS	1.01±0.19	1.06±0.17	1.17±0.27	1.35±0.31
TransCal	1.01±0.19	1.06±0.17	1.17±0.27	1.35±0.31
Cluster NN	1.01±0.19	1.06±0.17	1.17±0.28	1.36±0.32
Cluster LR	4.12±0.4	3.99±0.32	4.0±0.28	4.05±0.36
Cluster En.	2.11±0.12	2.11±0.14	2.21±0.23	2.36±0.34
CaliGen	**0.99±0.17**	1.04±0.16	1.15±0.26	**1.32±0.29**
CaliGen TS	1.01±0.18	1.06±0.16	1.17±0.25	1.37±0.3
CaliGen En.	0.99±0.18	**1.03±0.16**	**1.14±0.26**	1.33±0.3
Calibrated only on Calibration Domains				
HB-TL	1.5±0.29	1.57±0.28	1.79±0.44	2.19±0.64
TS	**1.01±0.19**	**1.06±0.17**	**1.17±0.27**	**1.35±0.31**
CPCS	**1.01±0.19**	**1.06±0.17**	**1.17±0.27**	**1.35±0.31**
TransCal	1.04±0.18	1.08±0.17	1.19±0.27	1.36±0.3
Cluster NN	1.06±0.19	1.11±0.17	1.22±0.26	1.4±0.3
Cluster LR	2.7±0.3	2.6±0.23	2.57±0.16	2.64±0.18
Cluster En.	1.68±0.07	1.67±0.07	1.72±0.16	1.86±0.24
CaliGen	1.14±0.2	1.2±0.18	1.3±0.26	1.52±0.33
CaliGen TS	1.33±0.24	1.4±0.22	1.51±0.27	1.8±0.4
CaliGen En.	1.03±0.19	1.08±0.17	1.19±0.26	1.39±0.31

Table 14: Cross-entropy loss for model trained on original images and 4 filters with corruption level 1, where column Train Filter and Calib. Filter represents the source and calibration domains while Test Filter (1) represents the rest of the test filter of severity of corruption 1 and All Filter (i) represents all the filters with severity of corruption i

Method	All Filter (3)	All Filter (4)	All Filter (5)	Average
Uncalibrated	2.01±0.59	2.36±0.8	2.64±0.91	1.75±0.61
Source Only (TS)	1.66±0.42	1.9±0.52	2.09±0.57	1.46±0.3
Oracle (TS)	*1.57±0.37*	*1.73±0.38*	*1.85±0.35*	*1.39±0.19*
Calibrated on Source and Calibration Domains				
HB-TL	2.29±0.7	2.74±1.02	3.09±1.29	2.02±0.92
TS	1.63±0.41	1.86±0.49	2.04±0.53	1.45±0.27
CPCS	1.62±0.4	1.84±0.47	2.02±0.51	1.44±0.26
TransCal	1.62±0.4	1.84±0.48	2.02±0.51	1.44±0.26
Cluster NN	1.66±0.42	1.89±0.53	2.08±0.58	1.46±0.3
Cluster LR	4.17±0.63	4.38±0.93	4.46±1.11	4.17±0.45
Cluster En.	2.59±0.52	2.85±0.7	2.99±0.8	2.46±0.34
CaliGen	**1.58±0.38**	**1.79±0.46**	**1.96±0.49**	**1.4±0.24**
CaliGen TS	1.65±0.42	1.9±0.53	2.09±0.59	1.46±0.3
CaliGen En.	1.59±0.39	1.81±0.47	1.98±0.5	1.41±0.25
Calibrated only on Calibration Domains				
HB-TL	2.79±0.92	3.39±1.33	3.9±1.68	2.45±1.64
TS	1.61±0.39	1.83±0.46	2.0±0.49	1.43±0.25
CPCS	1.61±0.38	1.82±0.45	1.98±0.48	1.43±0.24
TransCal	**1.59±0.36**	**1.79±0.41**	**1.94±0.43**	**1.43±0.21**
Cluster NN	1.7±0.41	1.94±0.53	2.13±0.6	1.51±0.3
Cluster LR	2.69±0.31	2.82±0.45	2.9±0.61	2.7±0.14
Cluster En.	2.05±0.37	2.26±0.47	2.37±0.57	1.94±0.18
CaliGen	1.81±0.46	2.01±0.54	2.15±0.54	1.59±0.29
CaliGen TS	2.18±0.64	2.44±0.78	2.61±0.8	1.89±0.52
CaliGen En.	1.67±0.42	1.87±0.49	2.02±0.51	1.46±0.26

Table 15: Cross-entropy loss evaluated on target domains of Office-Home dataset and averaged by target domains

Method	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated	in	4.64±0.62	4.65±0.11	3.37±0.55	3.04±0.77	3.93±0.92
TS	Source Only	3.59±0.33	3.65±0.22	2.69±0.36	2.5±0.35	3.11±0.61
TS	*Oracle*	*3.32±0.22*	*3.22±0.11*	*2.61±0.3*	*2.44±0.31*	*2.9±0.45*
HB-TL	Source and Calibration Domains	4.85±0.34	4.3±0.17	3.23±0.51	2.94±0.51	3.83±0.88
TS		3.37±0.26	3.38±0.12	2.63±0.27	2.47±0.28	2.96±0.48
CPCS		3.39±0.26	3.37±0.06	2.62±0.29	2.48±0.28	2.97±0.48
TransCal		4.16±0.88	3.81±0.77	2.91±0.63	2.62±0.15	3.37±0.92
Cluster NN		3.36±0.23	3.43±0.1	2.64±0.27	2.46±0.28	2.97±0.49
Cluster LR		3.35±0.22	3.46±0.11	2.65±0.29	2.45±0.27	2.98±0.49
Cluster En.		3.36±0.24	3.41±0.09	2.63±0.27	2.45±0.28	2.96±0.49
CaliGen		**3.11±0.1**	3.21±0.12	**2.33±0.08**	**2.07±0.22**	2.68±0.51
CaliGen TS		3.94±0.06	4.27±0.08	2.8±0.17	2.35±0.28	3.34±0.81
CaliGen En.		3.11±0.18	**3.15±0.09**	2.34±0.16	2.14±0.26	**2.68±0.48**
HB-TL	Calibration Domains only	4.33±0.39	4.41±0.27	3.24±0.43	2.91±0.59	3.72±0.79
TS		3.35±0.24	3.32±0.12	2.67±0.22	2.53±0.24	2.97±0.43
CPCS		3.35±0.24	3.3±0.07	2.67±0.23	2.57±0.2	2.97±0.41
TransCal		3.71±0.62	3.79±0.64	2.73±0.42	2.66±0.42	3.22±0.75
Cluster NN		3.34±0.24	3.37±0.12	2.66±0.23	2.5±0.26	2.97±0.45
Cluster LR		3.33±0.23	3.49±0.05	2.67±0.26	2.5±0.28	3.0±0.48
Cluster En.		3.33±0.24	3.37±0.08	2.65±0.23	2.49±0.26	2.96±0.45
CaliGen		3.18±0.06	3.2±0.13	2.43±0.16	**2.22±0.21**	2.76±0.46
CaliGen TS		3.95±0.27	4.33±0.22	2.82±0.24	2.41±0.29	3.38±0.83
CaliGen En.		**3.09±0.1**	**3.1±0.09**	**2.39±0.05**	2.24±0.23	**2.7±0.41**

Table 16: Cross-entropy loss evaluated on target domains of Office-Home dataset and averaged by target domains when classifier trained using EfficientNet v2 B0

Method	Calibrated on	Art	Clipart	Product	RealWorld	Average
Uncalibrated		4.44±0.42	3.97±0.76	3.05±0.34	2.77±0.61	3.56±0.88
TS	Source Only	3.24±0.17	3.05±0.25	2.4±0.09	2.15±0.23	2.71±0.49
TS	Oracle	3.08±0.05	2.86±0.17	2.37±0.06	2.11±0.24	2.61±0.41
HB-TL	Source and Calibration Domains	4.45±0.7	3.92±0.77	2.88±0.35	2.71±0.55	3.49±0.95
TS		3.11±0.07	2.9±0.18	2.37±0.06	2.12±0.23	2.63±0.43
CPCS		3.1±0.07	2.89±0.18	2.37±0.06	2.12±0.23	2.62±0.42
TransCal		3.9±0.78	3.49±1.01	2.59±0.39	2.43±0.52	3.1±0.94
Cluster NN		3.11±0.07	2.91±0.19	2.37±0.06	2.11±0.23	2.63±0.43
Cluster LR		3.11±0.07	2.92±0.19	2.36±0.05	2.11±0.23	2.63±0.43
Cluster En.		3.11±0.07	2.91±0.19	2.36±0.06	2.11±0.23	2.62±0.43
CaliGen		2.89±0.05	**2.65±0.28**	**2.07±0.19**	**1.77±0.28**	**2.35±0.5**
CaliGen TS		3.44±0.1	3.26±0.22	2.35±0.17	2.02±0.33	2.77±0.64
CaliGen En.		**2.88±0.05**	2.67±0.24	2.11±0.1	1.83±0.26	2.37±0.46
HB-TL	Calibration Domains only	4.25±0.55	3.89±0.8	2.86±0.29	2.71±0.51	3.43±0.87
TS		3.1±0.07	2.88±0.17	2.39±0.03	2.16±0.22	2.63±0.4
CPCS		3.09±0.06	2.87±0.17	2.4±0.02	2.16±0.22	2.63±0.39
TransCal		3.41±0.44	3.39±0.93	2.37±0.06	2.27±0.15	2.86±0.75
Cluster NN		3.1±0.07	2.88±0.18	2.38±0.03	2.15±0.22	2.63±0.41
Cluster LR		3.09±0.06	2.9±0.18	2.39±0.02	2.15±0.22	2.63±0.41
Cluster En.		3.09±0.07	2.88±0.18	2.38±0.03	2.15±0.22	2.63±0.41
CaliGen		2.92±0.14	2.68±0.35	2.21±0.31	1.91±0.28	2.43±0.48
CaliGen TS		3.55±0.15	3.48±0.5	2.51±0.27	2.07±0.35	2.9±0.72
CaliGen En.		**2.87±0.04**	**2.63±0.26**	**2.17±0.15**	**1.9±0.24**	**2.39±0.42**

Table 17: Cross-entropy loss evaluated on target domains of DomainNet dataset and averaged by target domains

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
Uncalibrated	3.61±0.5	7.05±0.65	4.86±0.85	6.85±1.29	3.78±0.77	4.5±0.33	5.11±1.58
Source Only (TS)	3.44±0.48	6.49±0.52	4.51±0.81	6.06±0.94	3.56±0.75	4.16±0.39	4.7±1.35
Oracle (TS)	*3.28±0.38*	*5.32±0.18*	*4.11±0.57*	*5.07±0.29*	*3.4±0.58*	*3.9±0.31*	*4.18±0.88*
Oracle (TS)	*3.28±0.1*	*5.32±0.06*	*4.11±0.09*	*5.07±0.06*	*3.4±0.08*	*3.9±0.1*	*4.18±0.78*
Calibrated on Source and Calibration Domains							
HB-TL	3.79±0.46	7.04±0.73	4.99±0.37	6.72±0.51	3.9±0.62	4.79±0.4	5.2±1.37
TS	3.31±0.38	5.76±0.38	4.17±0.64	5.55±0.72	3.46±0.56	3.92±0.33	4.36±1.09
CPCS	3.31±0.4	5.91±0.46	4.23±0.69	5.69±0.79	3.46±0.64	3.95±0.35	4.42±1.17
TransCal	3.32±0.39	6.3±0.53	4.34±0.68	5.8±0.86	3.45±0.63	4.02±0.35	4.54±1.28
Cluster NN	3.32±0.36	5.78±0.22	4.16±0.53	5.58±0.54	3.42±0.54	3.97±0.31	4.37±1.06
Cluster LR	3.33±0.35	5.91±0.42	4.22±0.47	5.8±0.57	3.51±0.46	3.96±0.29	4.45±1.12
Cluster En.	3.29±0.36	5.79±0.29	4.15±0.54	5.61±0.5	3.42±0.52	3.92±0.32	4.36±1.08
CaliGen	3.09±0.25	5.59±0.1	3.84±0.31	5.33±0.19	3.25±0.34	3.67±0.23	4.13±1.01
CaliGen TS	3.13±0.26	6.29±0.18	4.08±0.24	5.97±0.23	3.28±0.3	3.87±0.22	4.44±1.27
CaliGen En.	**3.0±0.31**	**5.47±0.23**	**3.79±0.46**	**5.19±0.39**	**3.16±0.44**	**3.59±0.29**	**4.03±1.02**
Calibrated only on Calibration Domains							
HB-TL	4.38±0.51	7.15±0.73	5.34±0.37	6.09±0.27	4.65±0.49	4.53±0.34	5.36±1.1
TS	3.41±0.36	5.56±0.3	4.15±0.56	5.35±0.54	3.56±0.49	3.96±0.31	4.33±0.94
CPCS	3.92±0.46	5.87±0.2	4.84±0.39	5.34±0.35	4.26±0.48	4.07±0.34	4.72±0.8
TransCal	3.9±0.48	6.0±0.13	4.87±0.39	5.39±0.46	4.27±0.48	4.05±0.36	4.75±0.86
Cluster NN	3.34±0.33	5.62±0.25	4.12±0.55	5.44±0.49	3.47±0.51	3.94±0.29	4.32±0.99
Cluster LR	3.6±0.7	5.71±0.31	4.22±0.53	5.84±0.5	4.57±1.51	4.06±0.38	4.67±1.13
Cluster En.	3.45±0.46	5.62±0.27	4.13±0.56	5.5±0.43	3.96±0.84	3.96±0.32	4.44±0.97
CaliGen	3.43±0.3	5.64±0.09	4.06±0.2	5.37±0.17	3.52±0.2	3.89±0.19	4.32±0.89
CaliGen TS	3.53±0.32	6.54±0.21	4.35±0.2	6.24±0.25	3.58±0.19	4.18±0.19	4.74±1.23
CaliGen En.	**3.17±0.3**	**5.37±0.2**	**3.86±0.35**	**5.09±0.29**	**3.31±0.35**	**3.68±0.25**	**4.08±0.9**

Table 18: Error % for model trained on 4 filters of corruption level 1 and calibration domains include training domains for (in) and not included for (out)

Method	Train Filter	Calib. Filter	Test Filter (1)	All Filter (2)
Uncalibrated	33.73±7.25	35.73±6.58	39.87±10.67	46.39±12.03
Calibrated on Source and Calibration Domains				
HB-TL	34.02±7.01	35.96±6.38	40.33±10.29	46.92±11.79
Cluster LR	46.32±4.14	47.25±4.14	49.85±7.97	54.62±9.45
Cluster En.	39.85±6.09	41.4±5.77	45.11±9.55	50.9±11.19
CaliGen	**33.05±6.46**	**34.82±5.9**	**38.85±9.68**	**45.26±11.21**
CaliGen En.	33.08±6.88	34.95±6.31	39.02±10.16	45.55±11.6
Calibrated only on Calibration Domains				
HB-TL	**34.14±7.2**	**36.15±6.51**	**40.19±10.54**	**46.73±11.91**
Cluster LR	39.75±4.73	40.99±4.42	44.37±8.89	49.67±10.28
Cluster En.	36.91±6.05	38.55±5.56	42.29±9.87	48.31±11.29
CaliGen	36.5±7.38	38.57±6.65	42.3±10.16	49.14±11.78
CaliGen En.	34.29±7.26	36.29±6.6	40.4±10.36	47.0±11.93

Table 19: Error % for model trained on 4 filters of corruption level 1 and calibration domains include training domains for (in) and not included for (out)

Method	All Filter (3)	All Filter (4)	All Filter (5)	Average
Uncalibrated	55.14±14.29	61.61±15.01	66.42±14.1	48.41±16.82
Calibrated on Source and Calibration Domains				
HB-TL	55.64±14.19	62.08±14.72	66.78±14.03	48.82±16.71
Cluster LR	61.02±11.96	66.28±13.02	69.95±12.45	56.47±13.03
Cluster En.	58.62±13.66	64.55±14.38	68.74±13.46	52.74±15.34
CaliGen	**53.7±13.56**	**59.98±14.32**	**64.92±13.78**	**47.22±16.13**
CaliGen En.	54.19±14.0	60.59±14.79	65.63±14.04	47.57±16.57
Calibrated only on Calibration Domains				
HB-TL	55.42±14.08	61.81±14.74	66.67±13.93	48.73±16.66
Cluster LR	57.29±12.68	63.14±13.61	67.47±13.09	51.81±14.49
Cluster En.	56.42±13.69	62.55±14.4	67.1±13.61	50.3±15.71
CaliGen	57.07±13.73	62.56±13.91	67.13±13.1	50.47±15.89
CaliGen En.	**55.42±14.03**	**61.42±14.62**	**66.2±13.76**	**48.72±16.44**

Table 20: Error % on Office-Home averaged by target domain where calibration domains (in) include source domain, (out) does not include source domain

Method	Art	Clipart	Product	RealWorld	Average
Uncalibrated	78.0±4.67	75.68±2.31	63.32±6.38	58.8±6.68	68.95±9.68
HB-TL (in)	77.68±4.94	75.08±1.35	63.0±6.56	58.29±6.33	68.51±9.64
CaliGen (in)	**72.71±2.4**	**71.94±0.17**	**57.97±1.69**	**51.34±5.36**	**63.49±9.64**
CaliGen En. (in)	73.76±3.93	72.45±1.5	58.75±4.07	52.42±6.23	64.35±10.01
HB-TL (out)	77.8±4.83	75.07±1.18	62.92±6.17	58.36±6.35	68.54±9.57
CaliGen (out)	74.57±1.41	73.54±1.12	60.31±3.75	54.15±5.21	65.64±9.31
CaliGen En. (out)	**73.36±2.19**	**72.25±1.44**	**58.5±2.54**	**52.59±6.09**	**64.18±9.57**

Table 21: Error % on Office-Home averaged by target domain where calibration domains (in) include source domain, (out) does not include source domain. The classifier was trained using EfficientNet V2 B0

Method	Art	Clipart	Product	RealWorld	Average
Uncalibrated	72.67±2.09	66.62±2.8	56.86±2.56	51.77±5.01	61.98±8.8
HB-TL (in)	72.39±2.18	66.43±3.16	56.23±2.61	51.23±5.07	61.57±8.99
CaliGen (in)	**68.33±0.72**	**62.58±6.0**	**51.12±3.48**	**44.9±6.45**	**56.74±10.37**
CaliGen En. (in)	68.9±1.19	63.75±4.59	51.9±1.33	46.0±6.05	57.64±9.92
HB-TL (out)	72.31±2.06	66.24±3.48	55.95±2.42	51.19±5.07	61.42±9.01
CaliGen (out)	68.33±2.83	**63.51±7.28**	54.46±6.14	46.84±6.6	58.28±10.2
CaliGen En. (out)	**68.23±0.35**	63.56±4.59	**52.25±0.81**	**46.37±5.55**	**57.6±9.43**

Table 22: Error % on DomainNet averaged by target domain where calibration domains (in) include source domain, (out) does not include source domain

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
Uncalibrated	65.0±6.23	90.22±2.53	76.9±8.6	92.48±3.22	68.99±8.51	73.96±4.97	77.93±11.93
CaliGen	63.86±4.66	**88.36±2.15**	73.9±5.11	91.32±1.75	68.98±5.5	71.57±3.9	76.33±10.86
CaliGen En.	**60.82±5.97**	88.41±2.55	**73.14±7.54**	**91.24±2.9**	**66.12±7.71**	**70.04±4.9**	**74.96±12.53**
CaliGen	70.29±5.29	89.99±1.18	78.36±3.12	92.06±1.3	73.91±3.05	75.62±3.51	80.04±8.77
CaliGen En.	**61.96±5.79**	**88.52±2.3**	**73.6±6.42**	**91.49±2.65**	**66.32±6.38**	**70.67±4.7**	**75.43±12.04**

Table 23: Temperature value for model trained on original images and 4 filters with corruption level 1, where column Train Filter and Calib. Filter represents the source and calibration domains while Test Filter (1) represents the rest of the test filter of severity of corruption 1 and All Filter (i) represents all the filters with severity of corruption i

Method	Train Filter	Calib Filter	Test Filter (1)	All Filter (2)
TS (Source Only)	1.58±0.0	1.58±0.0	1.58±0.0	1.58±0.0
TS (Oracle)	1.64±0.11	1.66±0.12	1.73±0.07	1.88±0.23
CPCS	1.77±0.0	1.77±0.0	1.77±0.0	1.77±0.0
TransCal	1.76±0.0	1.76±0.0	1.76±0.0	1.76±0.0
TS	1.7±0.0	1.7±0.0	1.7±0.0	1.7±0.0
Cluster NN	2.86±0.08	2.85±0.09	2.84±0.07	2.81±0.09
Cluster LR	-0.1±0.75	0.03±0.8	0.11±0.67	0.11±0.75

Table 24: Temperature value for model trained on original images and 4 filters with corruption level 1, where column Train Filter and Calib. Filter represents the source and calibration domains while Test Filter (1) represents the rest of the test filter of severity of corruption 1 and All Filter (i) represents all the filters with severity of corruption i

Method	All Filter (3)	All Filter (4)	All Filter (5)	Average
TS (Source Only)	1.58±0.0	1.58±0.0	1.58±0.0	1.58±0.0
TS (Oracle)	2.24±0.51	2.66±0.92	3.18±1.43	2.14±0.87
CPCS	1.77±0.0	1.77±0.0	1.77±0.0	1.77±0.0
TransCal	1.76±0.0	1.76±0.0	1.76±0.0	1.76±0.0
TS	1.7±0.0	1.7±0.0	1.7±0.0	1.7±0.0
Cluster NN	2.77±0.1	2.76±0.09	2.74±0.09	2.8±0.1
Cluster LR	-0.01±0.83	-0.1±0.95	-0.17±1.0	-0.02±0.83

Table 25: Temperature value evaluated on target domains of Office-Home dataset and averaged by target domains, when classifier is trained using EfficientNet V2 B0

Method	Art	Clipart	Product	RealWorld	Average
TS (Source Only)	1.82±0.18	1.6±0.24	1.72±0.34	1.67±0.33	1.7±0.29
TS (Oracle)	2.55±0.26	2.26±0.47	1.99±0.33	1.95±0.42	2.19±0.45
Calibrated on Source and Calibration Domains					
CPCS	2.24±0.12	1.94±0.34	2.06±0.41	2.12±0.41	2.09±0.36
TransCal	1.38±0.46	1.42±0.48	1.72±0.47	1.54±0.43	1.52±0.48
TS	2.17±0.1	1.88±0.33	2.03±0.4	2.09±0.41	2.04±0.35
Cluster NN	2.16±0.08	1.85±0.3	2.02±0.38	2.09±0.42	2.03±0.34
Cluster LR	2.17±0.1	1.84±0.28	2.01±0.37	2.1±0.43	2.03±0.34
Calibrated on Calibration Domains Only					
CPCS	2.4±0.15	2.09±0.45	2.27±0.55	2.35±0.56	2.28±0.47
TransCal	1.7±0.36	1.72±0.58	2.0±0.39	1.82±0.42	1.81±0.46
TS	2.31±0.14	2.0±0.41	2.21±0.5	2.3±0.54	2.21±0.44
Cluster NN	2.31±0.12	1.96±0.37	2.19±0.48	2.28±0.53	2.18±0.43
Cluster LR	2.32±0.12	1.91±0.34	2.13±0.45	2.27±0.49	2.16±0.41

Table 26: Temperature value evaluated on target domains of DomainNet dataset and averaged by target domains, when calibration domains include source domains

Method	clipart	infograph	painting	quickdraw	real	sketch	Average
TS (Source Only)	1.14±0.09	1.16±0.09	1.18±0.12	1.22±0.09	1.2±0.1	1.2±0.1	1.18±0.1
TS (Oracle)	1.51±0.15	2.95±0.62	1.87±0.29	2.7±0.77	1.58±0.2	1.76±0.12	2.06±0.71
CPCS	1.48±0.23	1.47±0.23	1.49±0.2	1.41±0.13	1.55±0.22	1.57±0.25	1.49±0.22
TransCal	1.35±0.22	1.27±0.24	1.32±0.15	1.35±0.13	1.42±0.16	1.4±0.22	1.35±0.2
TS	1.62±0.23	1.58±0.23	1.63±0.21	1.51±0.14	1.74±0.24	1.7±0.25	1.63±0.23
Cluster NN	1.72±0.14	1.57±0.23	1.86±0.3	1.51±0.19	1.85±0.19	1.83±0.17	1.72±0.25
Cluster LR	1.71±0.19	1.59±0.34	1.94±0.4	1.32±0.34	1.83±0.28	1.88±0.17	1.71±0.36

Table 27: Calibration performance (ECE %) averaged by target and source domains of Office-Home dataset while ResNet trained on all combinations of 3 domains

Method	Averaged by	Art	Clipart	Product	RealWorld	Average
Uncalibrated	Target	30.35±5.35	40.05±10.54	19.94±4.72	23.59±6.04	28.48±10.4
TS	Domains	10.16±0.85	18.88±2.27	6.61±1.68	6.86±0.79	10.63±5.19
Uncalibrated	Source	0.43±0.36	0.47±0.04	0.74±1.11	0.27±0.05	0.48±0.61
TS	Domains	6.12±6.33	2.1±0.56	5.27±8.3	1.19±0.99	3.67±5.65