PCL: Proxy-based Contrastive Learning for Domain Generalization Supplementary Material

Xufeng Yao†, Yang Bai†, Xinyun Zhang†, Yuechen Zhang†, Qi Sun†, Ran Chen†, Ruiyu Li♯, Bei Yu†
†The Chinese University of Hong Kong ♯SmartMore
{xfyao,byu}@cse.cuhk.edu.hk

1. Proof of Equations

Proof of Equation (4).

We can prove the Equation (4) utilizing the log-sum-exp inequalities [12]. We define the Log-Sum-Exp function as $LSE(.)$, then we can have the following bounds:

$$\begin{aligned}\max\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} &\leq LSE(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) \\ &\leq \max\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} + \log(n)\end{aligned} \tag{1}$$

The inequality holds if and only if $n = 1$. We can make the bound tighter by multiplying a scale factor α. Then:

$$\begin{aligned}\max\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} &\leq \frac{1}{\alpha} LSE(\alpha\boldsymbol{x}_1, \alpha\boldsymbol{x}_2, \cdots, \alpha\boldsymbol{x}_n) \\ &\leq \max\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} + \frac{\log(n)}{\alpha}\end{aligned} \tag{2}$$

When $\alpha \to \infty$, then equation holds. Then as for equation 4, we have:

$$\begin{aligned}\mathcal{L}_{CL} &= \lim_{\alpha\to\infty} \frac{1}{\alpha} - \log \frac{\exp(\alpha s_p)}{\exp(\alpha s_p) + \sum_{j=1}^{\alpha} \exp(\alpha s_n^j)} \\ &= \lim_{\alpha\to\infty} \frac{1}{\alpha} \log(1 + \sum_{j=1}^{\alpha} \exp(\alpha s_n^j - s_p)) \\ &\geq \lim_{\alpha\to\infty} \log(\sum_{j=1}^{\alpha} \exp(s_n^j - s_p)) \\ &= \lim_{\alpha\to\infty} \frac{1}{\alpha} LSE(\alpha s_n^j - \alpha s_p) \\ &= \max[s_n^j - s_p]\end{aligned} \tag{3}$$

Proof of Equation (5).

If s_i is the positive score, then the derivative can be given by:

$$\begin{aligned}\frac{\partial \mathcal{L}}{s_i} &= -\frac{\sum_{j=1}^{C} \exp(s_j)}{\exp(s_i)} \cdot \\ &\quad \frac{\exp(s_i)\sum_{j=1}^{C}\exp(s_j) - (\exp(s_i))^2}{(\sum_{j=1}^{C}\exp(s_j))^2} \\ &= -\frac{\sum_{j=1}^{C}\exp(s_j) - \exp(s_i)}{\sum_{j=1}^{C}\exp(s_j)} \\ &= \frac{\exp(s_i)}{\sum_{j=1}^{C}\exp(s_j)} - 1 \\ &= p_i - 1.\end{aligned} \tag{4}$$

If s_j is the negative score, the the derivative can be given by:

$$\begin{aligned}\frac{\partial \mathcal{L}}{s_j} &= -\frac{\sum_{j=1}^{C}\exp(s_j)}{\exp(s_i)} \frac{-\exp(s_j)\exp(s_i)}{(\sum_{j=1}^{C}\exp(s_j))^2} \\ &= -\frac{-\exp(s_j)}{\sum_{j=1}^{C}\exp(s_j)} = \frac{\exp(s_j)}{\sum_{j=1}^{C}\exp(s_j)} \\ &= p_j.\end{aligned} \tag{5}$$

Since $\sum p_i = 1$, assume we have B pairs, then we have $\sum_{j=1}^{B-1} \frac{\partial \mathcal{L}}{s_j} = |\frac{\partial \mathcal{L}}{s_i}|$.

Intution behind Equation (1). We also provide the intuition behind Equation 1 here. The positive alignment loss function is not constructed casually. Based on the proof of Equation (4), when the scale factor $\alpha \to \infty$ we can have:

$$\begin{aligned}\mathcal{L}_{pos} &= \lim_{\alpha\to\infty} \frac{1}{\alpha} \log(1 + \sum \exp(-\boldsymbol{z}_i^\top \boldsymbol{z}_j \cdot \alpha)) \\ &\geq \lim_{\alpha\to\infty} \frac{1}{\alpha} \log(\sum \exp(-\boldsymbol{z}_i^\top \boldsymbol{z}_j \cdot \alpha)) \\ &= \min[\boldsymbol{z}_i^\top \boldsymbol{z}_j]\end{aligned} \tag{6}$$

where $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ represent the different samples that are sampled from the same class. In contrastive-based loss

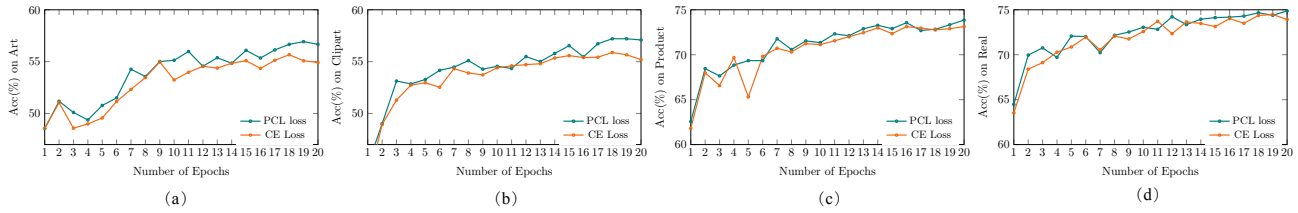


Figure 1. Comparison on different target domain on OfficeHome Benchmark

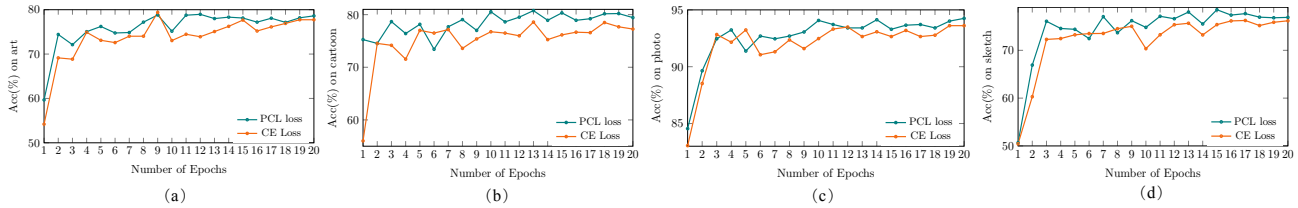


Figure 2. Comparsion on different target domain on PACS Benchmark

function, we attempt to conduct hard pair mining on controlling the scale factor by finding the most difficult negative pair. Similarly, we also want to conduct hard pair mining on positive pair. The above positive alignment loss provides one solution for hard positive pair mining.

2. More Experimental Analysis

Our code is mainly built on the open-source code of SWAD [4] including its training strategy.

Optimization Details. The network is optimized by Adam optimizer with a learning rate of 5e-5. All the input images are resized to 224 $\times$ 224. For all the datasets except domainnet, we train the model for 5000 steps. For the domainnet dataset, we train the model for 15000 steps. Note that this training setting is akin to the SWAD. We also follow the same HP searching strategy as SWAD did.

Data Augmentation Details. Data augmentation plays a vital role in domain generalization as a typical regularization method. Though there are many data augmentation methods such as Jigen [3] show a promising result on DG task. For a fair comparison, we only use the data augmentations contained in SWAD. We follow the data augmentation technicals in SWAD. We randomly cropped the images to retain between 70% and 100%. We randomly applied horizontal flipping and random color jittering with a magnitude of 0.3. We also randomly apply a Grayscale on the original image with 10% probabilities.

More Experimental Results. We also validate our algorithm on VLCS [5], which contains about 11K images with four domains. As shown in Table 1, our method does not surpass the state-of-the-art methods. We also report the reproduced SWAD results on VLCS (i.e., SWAD†). Our approach surpasses the reproduced results.

Table 1. Comparison with state-of-the-art methods on VLCS benchmark with ResNet-50 imagenet-pretrained model

Algorithm	C	L	S	V	Avg
GroupDRO [14]	97.3 $\pm$0.3	63.4 $\pm$0.9	69.5 $\pm$0.8	76.7 $\pm$0.7	76.7
RSC [7]	97.9	62.5	72.3	75.6	77.1
MLDG [9]	97.4	65.2	71.0	75.3	77.2
MTL [2]	97.8	64.3	71.5	75.3	77.2
ERM [16]	98.0	64.7	71.4	75.2	77.3
I-Mixup [17–19]	98.3	64.8	72.1	74.3	77.4
ERM [16]	97.7	64.3	73.4	74.6	77.5
MMD [10]	97.7	64.0	72.8	75.3	77.5
CDANN [10]	97.1	65.1	70.7	77.1	77.5
ARM [20]	98.7	63.6	71.3	76.7	77.6
SagNet [11]	97.9	64.5	71.4	77.5	77.8
Mixstyle [21]	98.6	64.5	72.6	75.7	77.9
VREx [8]	98.4	64.4	74.1	76.2	78.3
IRM [1]	98.6	64.9	73.4	77.3	78.6
DANN [6]	**99.0**	65.1	73.1	77.2	78.6
CORAL [15]	98.3	**66.1**	73.4	77.5	78.8
SWAD [4]	98.8	63.3	**75.3**	**79.2**	**79.1**
SWAD†	98.41	63.58	72.01	74.49	77.12
Ours	**99.02**	63.57	73.75	75.58	77.98

Convergence Speed. We also conduct an experiment on analyzing the convergence speed of our method. We use OfficeHome and PACS datasets and test the convergence speed with a ResNet18 backbone with imagenet-pretrained model. We do not use SWA mechanism in our implementation.

As shown in Figures 1 and 2, in all domain generalization settings, the proposed PCL loss can converge faster than the softmax CE loss and achieve a better model performance.

3. Visualization results

We also use deep neural network interpretability methods in [13] to explain the our model's generalization ability, as shown in Figures 3 to 6, the Art, Clipart, Product and Real-World indicates the target domain correspondingly. Our model can capture the important part of the input

sample.

References

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

[2] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 22:2–1, 2021. 2

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2229–2238, 2019. 2

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[5] 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*, pages 1657–1664, 2013. 2

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

[7] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision (ECCV)*, pages 124–140. Springer, 2020. 2

[8] 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*, pages 5815–5826. PMLR, 2021. 2

[9] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[10] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2018. 2

[11] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 2

[12] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016. 1

[13] Utku Ozbulak. PyTorch CNN Visualizations. https://github.com/utkuozbulak/pytorch-cnn-visualizations, 2019. 2

[14] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 2

[15] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 2

[16] V Vapnik. Statistical learning theory new york. *NY: Wiley*, 1:2, 1998. 2

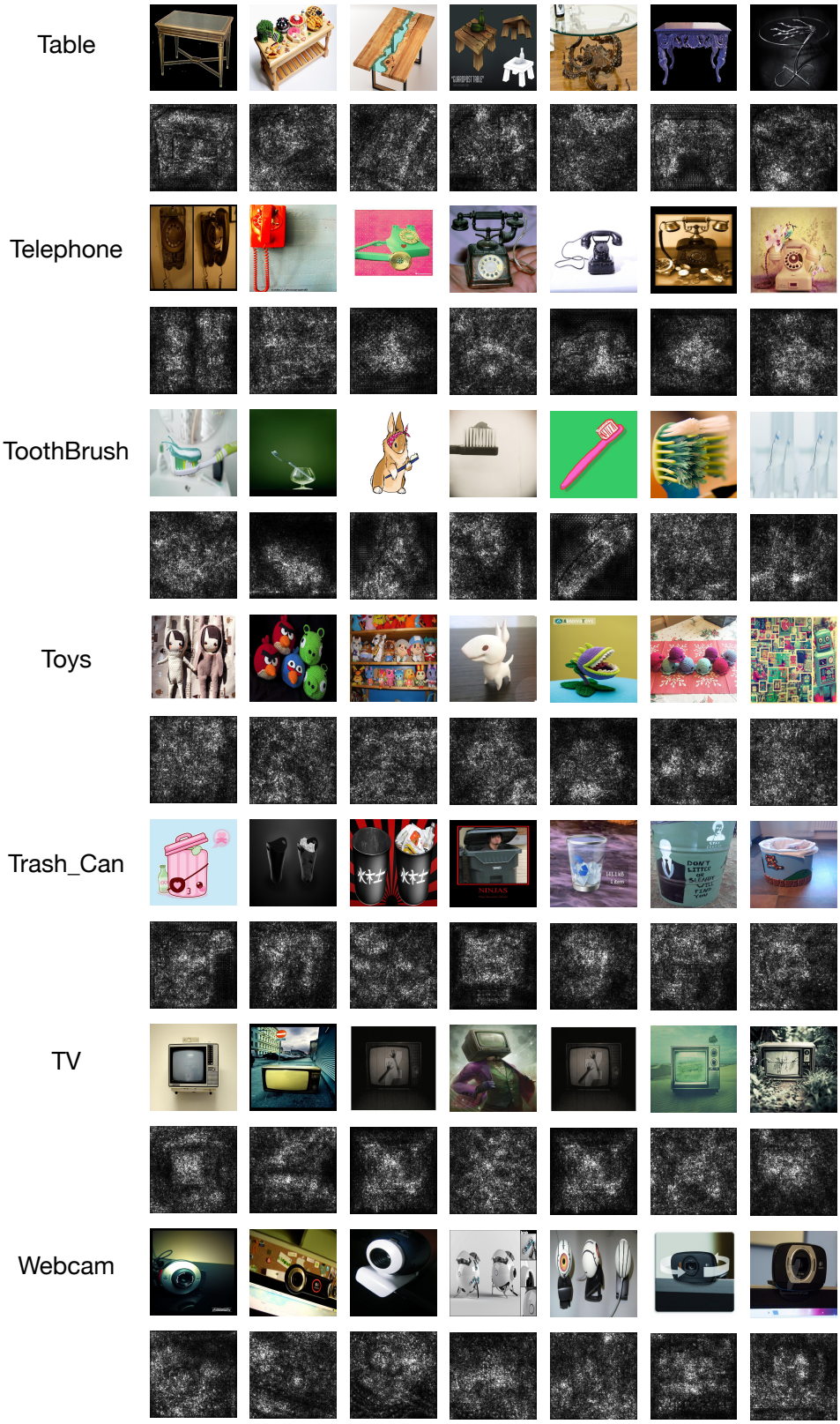
[17] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020. 2

[18] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020. 2

[19] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 2

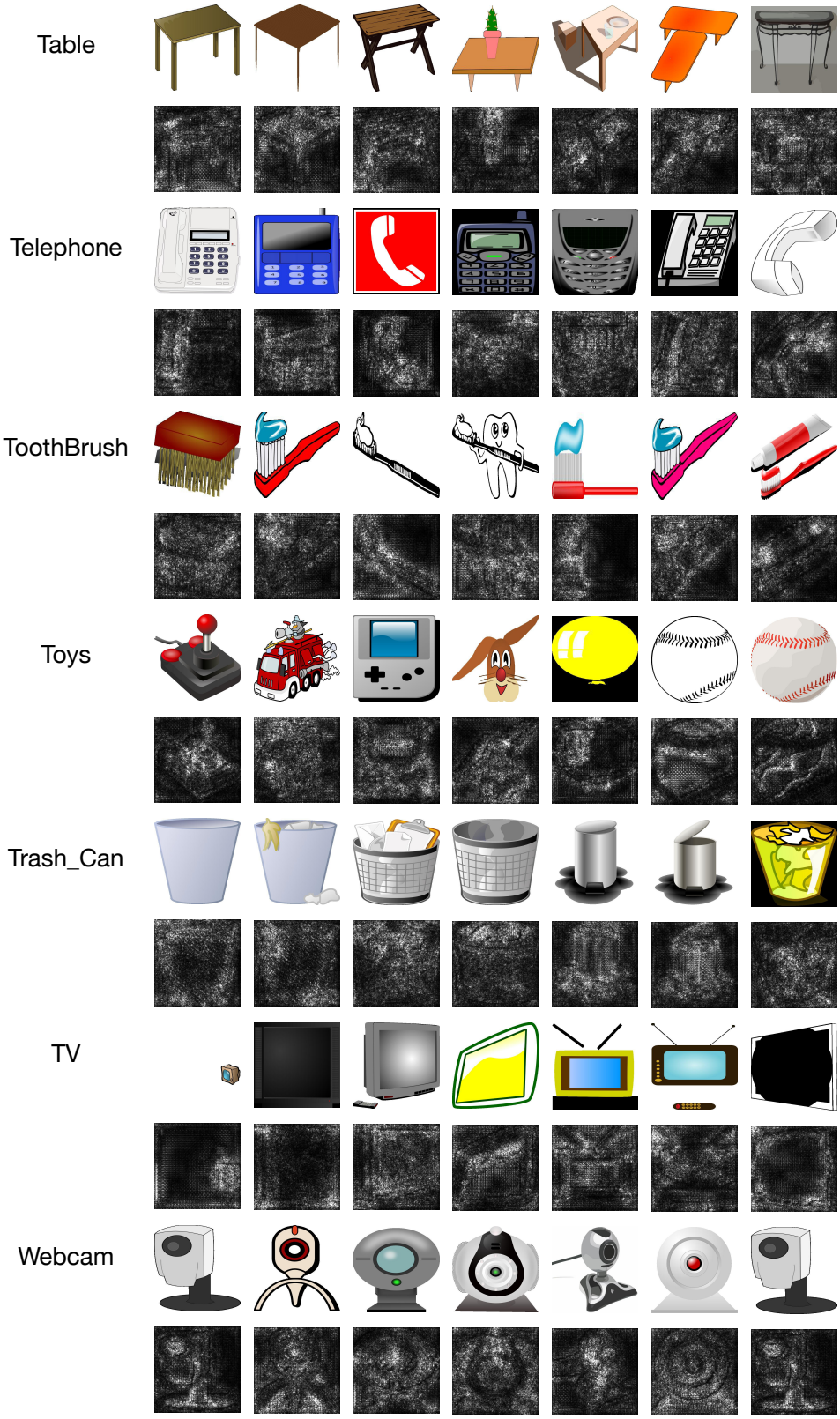
[20] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group distribution shift. *arXiv preprint arXiv:2007.02931*, 2020. 2

[21] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 2



(a) Art

Figure 3. Visualization results on OfficeHome Art



(b) Clipart

Figure 4. Visualization results on OfficeHome Clipart

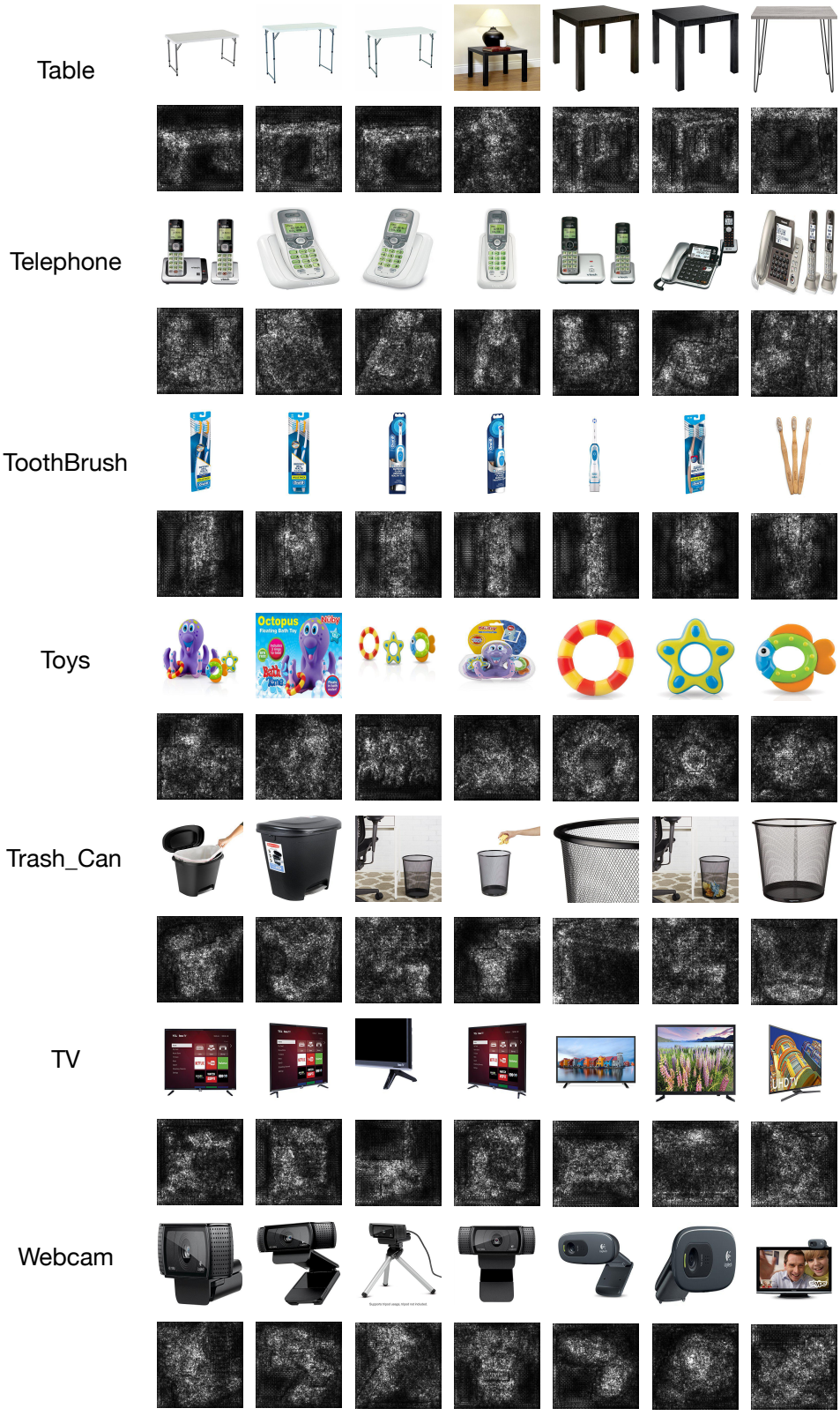
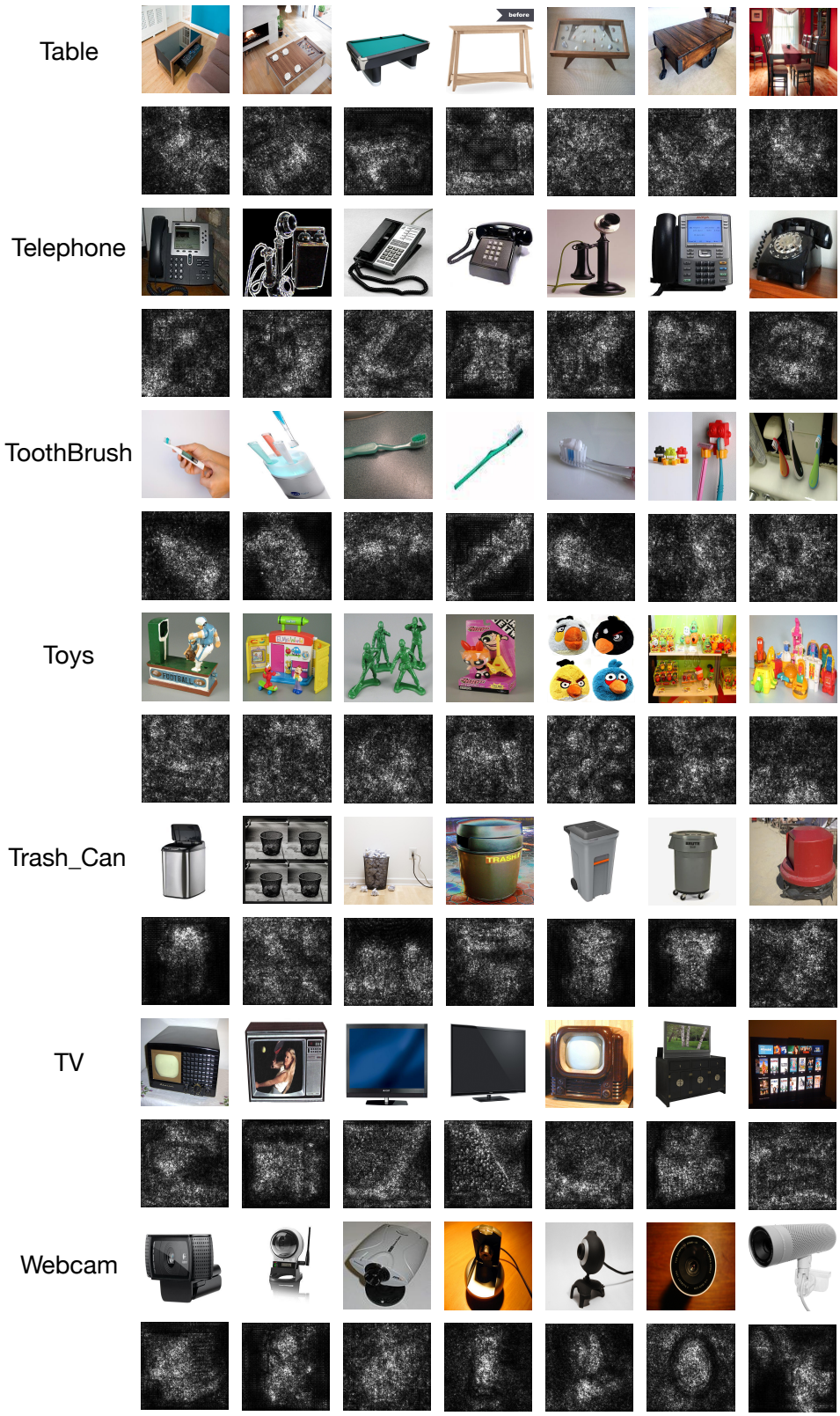


Figure 5. Visualization results on OfficeHome Product



(d) Real_World

Figure 6. Visualization results on OfficeHome Real-World