On Challenges in Unsupervised Domain Generalization

Vaasudev Narayanan cs20mtech11001@iith.ac.in
Department of Computer Science & Engineering, IIT Hyderabad

Aniket Anand Deshmukh aniketde@umich.edu
Microsoft, Mountain View, CA, USA

Urun Dogan urundogan@gmail.com
Microsoft, Mountain View, CA, USA

Vineeth N Balasubramanian vineethnb@cse.iith.ac.in
Department of Computer Science & Engineering, IIT Hyderabad

Abstract

Domain Generalization (DG) aims to learn a model from a labeled set of source domains which can generalize to an unseen target domain. Although an important stepping stone towards building general purpose models, the reliance of DG on labeled source data is a problem if we are to deploy scalable ML algorithms in the wild. We thus propose to study a novel and more challenging setting which shares the same goals as that of DG, but without source labels. We name this setting as Unsupervised Domain Generalization (UDG), where the objective is to learn a model from an unlabeled set of source domains that can *semantically cluster* images in an unseen target domain. We investigate the challenges involved in solving UDG as well as potential methods to address the same. Our experiments indicate that learning a generalizable feature representation using self-supervision is a strong baseline for UDG, even outperforming sophisticated methods explicitly designed to address domain shift and clustering.

Keywords: Unsupervised Domain Generalization, Clustering

1. Introduction

There is a growing need for generalizable models that learn with limited supervision. Safety-critical systems such as self-driving cars require robust models that are invariant to changes between train and test distributions such as weather (Volk et al., 2019), illumination (Dai and Gool, 2018) and location Varma et al. (2019). To this end, considerable research has been conducted in domain generalization (DG) (Blanchard et al., 2021; Li et al., 2017; Carlucci et al., 2019; Ghifary et al., 2015; Li et al., 2018b; Muandet et al., 2013; Gulrajani and Lopez-Paz, 2020) in recent years. DG has been formulated as learning a model from a labeled set of source domains that can generalize to an unseen target domain. While an important step towards deploying machine learning algorithms in the wild, the necessity of labeled source data in the formulation of DG is a bottleneck. Data labeling can be a time-consuming process – and in certain niche applications like healthcare that require subject matter experts, it can quickly get infeasible, especially on multiple source domains. There is clearly tremendous incentive to build models which obviate the need of labeling images in the source domain (even if partially on some of the source domains) and can generalize to unseen target domains (Blanchard et al., 2021).

We thus propose in this work to study a new problem setting which we name Unsupervised Domain Generalization (UDG), where the objective is to learn a model from an unlabeled set of source domains such that it can *semantically cluster* images in an unseen target domain. A mathematical formulation of our problem setting is provided in Section 2.1. Despite the potential applicability of the problem setting in real-world scenarios, there surprisingly has not been a principled approach to study it. To the best of our knowledge, ours is the first such effort.

Considering the nature of the constraints in the UDG problem, we have preliminarily identified two key challenges to be addressed herein – *unsupervised learning under domain shift* and *cluster imbalance* (see Figure 1). Existing DG methods (which require labeled source domains) address the domain shift problem. We investigate in this work whether the ideas in existing DG solutions can be translated and adapted to tackling UDG. We expand on this in the next section. UDG is also related to Unsupervised Clustering under Domain Shift (UCDS) (Menapace et al., 2020), a slightly relaxed setting where in addition to unlabeled source images for, access to unlabeled images in the target domain is also allowed. Zhang et al. (2021) propose a setting that shares the same name as ours but assume availability of labeled source data for fine-tuning. Various related settings in domain shift literature are summarized in Table 1.

The second challenge that needs we recognize is that of cluster imbalance. An implicit assumption among contemporary deep clustering methods (Niu and Wang, 2021; Deshmukh et al., 2021; Li et al., 2021; Regatti et al., 2020; Van Gansbeke et al., 2020) that are designed for a single domain is that the clusters ought to be balanced. This is, in part, to avoid trivial (or empty) clusters, and also due to the nature of existing clustering benchmark datasets which are all balanced: CIFAR-10 & CIFAR-100 (Krizhevsky et al., 2009), STL-10 (Coates et al., 2011), ImageNet-10 (Howard, 2019) and ImageNet-Dogs (Deng et al., 2009). While important for making initial forays, this is too strong an assumption for real-world datasets. For example, Figure 1 illustrates the strong imbalance present inside each domain in the PACS (Li et al., 2017) dataset (this is one of the simpler DG datasets, to put this in context).

The main contributions of this proposal can be summarized as follows: (i) We introduce a new problem setting of Unsupervised Domain Generalization, where we are provided with an unlabeled set of source domains and we aim to learn a model that semantically clusters images on an unseen target domain; (ii) We outline the challenges in solving UDG – in particular, *unsupervised learning under domain shift* and *imbalanced clustering* and propose solutions to address these challenges as a first effort; and (iii) We conduct empirical evaluation on standard benchmark datasets in the domain shift literature and seek to establish a solid baseline for the UDG problem.

2. Addressing Challenges in UDG

In this section, we give a mathematical formulation for UDG, describe how solutions from existing literature on domain generalization and clustering can transfer to our problem setting, and propose a potential methodology to tackle UDG.

Table 1: Comparison with other related settings in literature. U: Unlabeled, PL: Partially Labeled, L: Labeled

Setting	Source			Target		
	U	PL	L	U	PL	L
Domain Adaptation (DA) (Wang and Deng, 2018)	×	×	✓	×	×	✓
Semi-supervised DA (Saito et al., 2019)	×	×	✓	×	✓	×
Unsupervised DA (Wilson and Cook, 2020)	×	×	✓	✓	×	×
Domain Generalization (Blanchard et al., 2021)	×	×	✓	×	×	×
CDS (Kim et al., 2020)	×	✓	×	✓	×	×
UCDS (Menapace et al., 2020)	✓	×	×	✓	×	×
UDG (Ours)	✓	×	×	×	×	×

2.1. Problem Formulation

Let $\mathcal{X} \in \mathcal{R}^{H \times W \times Ch}$ be the input space of images (where H and W are the height and width of the images, Ch denotes the number of channels), and $\mathcal{F} \in \mathcal{R}^d$ be the hidden representation space. We assume access to M source domains $\{D_i^s\}_{i=1}^M$ where the i^{th} source domain D_i^s contains N_i unlabeled instances $\{x_j^i\}_{j=1}^{N_i}$. Using $\{D_i^s\}_{i=1}^M$, we learn a feature extractor $f_\theta : \mathcal{X} \rightarrow \mathcal{F}$ and a cluster classifier $f_\phi : \mathcal{F} \rightarrow \mathcal{C}$, where $\mathcal{C} \in \{1, \dots C\}$ and C is the number of desired clusters. $f_\phi(f_\theta)$ can then be used to cluster images for an unseen target domain D^t. We call this setting following recent work (Niu and Wang, 2021; Deshmukh et al., 2021; Li et al., 2021; Regatti et al., 2020; Van Gansbeke et al., 2020) as *semantic clustering* due to the evaluation procedure at test time, which involves checking for correct assignment to class labels. We do not make any assumptions about the underlying distribution for either the source or the target domain. The number of classes are assumed to be the same across all domains for evaluation purposes.

2.2. Can Existing Solutions Transfer to UDG?

Based on preliminary studies, we identified two major challenges that we need to address to solve UDG - domain shift and cluster imbalance. In this section, we look at how existing methods in DG and clustering could help us in addressing these challenges.

Domain Generalization Multiple approaches have been considered to handle domain shift (Zhou et al., 2021) - domain alignment, meta-learning, data augmentation, self-supervision and learning disentangled representations. Based on the success of meta-learning methods in DG (Li et al., 2018a; Dou et al., 2019; Li et al., 2019b,a; Zhao et al., 2021; Choi et al., 2021; Du et al., 2020; Liu et al., 2020), we propose to examine it for our use case. The key idea behind using meta-learning for DG is that domain shift is simulated by dividing the source domains into meta-train and meta-test domains. The model is trained on the meta-source domains and optimized for performance on the meta-test domain. By training in this fashion across epochs, the model is expected to generalize to an unseen domain.

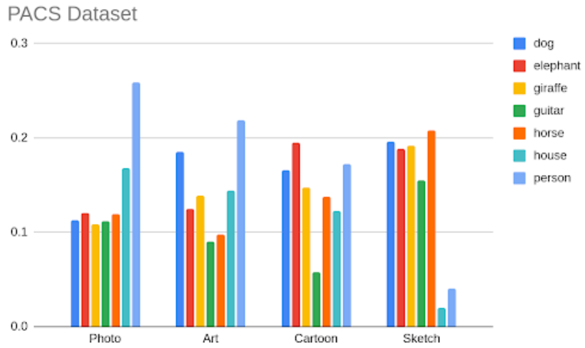


Figure 1: Imbalance across domains; Relative frequency for a particular class inside each domain

Clustering A plethora of deep learning-based methods have been proposed in clustering for datasets with no domain shift. The general norm in prevalent clustering methods is to train and test on the same dataset. UDG presents an additional challenge where a model is trained on one domain and tested on a completely different domain. Clustering methods broadly fall into two categories: (i) directly output a distribution over clusters, or (ii) learn a representation space that is amenable to being clustered using a simple technique like K-means. We plan to study methods from both categories (Tao et al., 2021; Van Gansbeke et al., 2020; Deshmukh et al., 2021; Niu and Wang, 2021; Han et al., 2020; Li et al., 2021).

Algorithm 1: Meta-Clustering for UDG

```
Input: $f_\theta$: Backbone network, $k$: Number of clusters
Output: $f_\theta^k$: Backbone network trained on $k$ clusters
$P(\mathcal{D})$: Distribution over domains
$f_\theta : \mathcal{R}^{H \times W \times Ch.} \to \mathcal{R}^d$
Clustering Head: $f_\phi : \mathcal{R}^d \to \mathcal{R}^k$
$f_\eta : f_\phi(f_\theta)$
while convergence criteria not satisfied do
    Sample meta-train domains $S_i \sim P(\mathcal{D})$
    Sample meta-test domain $T \sim P(\mathcal{D}) \mid T \neq S_i$
    forall $S_i$ do
        Sample a batch $B_i = \{x_p^{(i)}\}_{p=1}^{n_i} \sim S_i$
        Evaluate $\nabla_\eta L_{cluster}(f_\eta)$ on $B_i$
        Compute adapted parameters: $\eta_i^{'} := \eta - \alpha \nabla_\eta L_{cluster}(f_\eta)$
    end
    Sample a batch $B_T = \{x_j^T\}_{j=1}^{n_T} \sim T$
    Update $\eta := \eta - \beta \nabla_\eta \sum_{S_i} L_{cluster}(f_{\eta_i^{'}})$ using $B_T$
end
```

2.3. Curriculum Learning-based Meta-Learning Methodology for UDG

Hypothesis 1: Can existing methods to handle domain shift in DG be translated and adapted to tackle UDG?

Meta-learning, by design of 'learning to learn', provides a natural way to generalize across domains. Hence, on the back of the success of meta-learning methods in DG (Li et al., 2018b; Dou et al., 2019), we ask whether a meta-learning approach can be used for handling domain shift in UDG. As given in Algorithm 1, during training, we split the source domains into meta-source domains and meta-target domains, which are disjoint. We train a separate model for each meta-source domain and jointly optimize for domain shift by maximizing performance on a meta-target domain. We propose to experiment with and modify multiple state-of-the-art clustering objectives (Van Gansbeke et al., 2020; Niu and Wang, 2021; Tao et al., 2021; Han et al., 2020; Li et al., 2021; Deshmukh et al., 2021). Unlike prevalent deep clustering algorithms, we do not plan to add an entropy term to our objective as it encourages uniformly distributed clusters. Specific experiments are discussed below.

Algorithm 2: Curriculum Learning for Handling Cluster Imbalance

```
$\mathcal{C}$: No. of classes in the dataset
$c_i = [2, 4, 8, \dots, \mathcal{C}]$
Randomly initialized neural network $f_\theta : \mathcal{R}^{H \times W \times Ch.} \rightarrow \mathcal{R}^d$
Initialize $f_\theta^{(c_0)} \leftarrow f_\theta$
forall $c_i$ do
    $f_\theta^{(c_i)} \leftarrow MetaClusteringUDG(f_\theta^{(c_{i-1})}, c_i)$
end
```

Hypothesis 2: Can a curriculum learning strategy address cluster imbalance in UDG?

In order to address cluster imbalance, we propose to examine a curriculum based strategy, summarized in Algorithm 2 inspired from Dogan et al. (2020). We first run the the algorithm described in 1 with two clusters and train until convergence. Then, we train with four clusters with the model trained on two clusters as an initialization, converge it. We do this till the number of clusters is equal to the number of classes. Our hypothesis is that at the highest level of two clusters, the imbalance would be minimal. This is equivalent to clustering the data into two *superclasses.* Once the representations are refined at this level, we train our model with more clusters and eventually train it with C clusters. Thus at each level, by refining the representation for that level, we expect the model to counteract cluster imbalance.

3. Experimental Protocol & Planned Implementation

In this section, we discuss about (i) baseline methods for UDG, (ii) datasets & metrics we plan to use (iii) evaluation protocol and (iv) implementation details

3.1. Baselines

We aim to explore and establish multiple baselines for our problem setting.

- **Random + K-Means**: The simplest baseline is to get the representations from a randomly initialized convolutional neural network (CNN) (He et al., 2015) and perform K-Means clustering on those representations. The random baseline is a sensible lower bound.

- **SSL + K-Means**: A second baseline is to use a self-supervised (SSL) (Jing and Tian, 2020) method to train on a dataset formed by aggregating all the source domains and test it on the target domain.

- **Deep Clustering**: Following the previous two relatively simple baselines, we then propose to use a state-of-the-art deep clustering method which has been designed for a single domain (Van Gansbeke et al., 2020; Niu and Wang, 2021; Tao et al., 2021). We will follow the same evaluation strategy as that of the SSL baseline.

- **ACIDS w/o TA**: We will consider a modification to the UCDS setting from (Menapace et al., 2020). (Menapace et al., 2020) proposes a two-stage approach where a CNN is first trained on source domains and in the second stage is fine-tuned on the target domain. By discarding the target fine-tuning step, we obtain a strong baseline for UDG.

- **ImageNet + K-Means**: We extract representations from an ImageNet pre-trained ResNet18 model for the target domain and cluster them using K-Means.

3.2. Datasets & Metrics

Datasets We propose to perform our experiments on three datasets which are commonly used in the domain shift literature for evaluation - PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017) and Office31 (Saenko et al., 2010). PACS comprises 9,991 images divided in 7 classes across 4 different domains: Photo, Art, Cartoon and Sketch. The Office-Home dataset contains about 15,500 images across 4 domains: Product, Art, Clipart and Real World. Each domain is divided into 65 different classes. The Office31 dataset contains 4,110 images divided into 31 classes across 3 different domains: Amazon, DSLR, and Webcam. Each of these datasets contain visually disparate domains across a wide spectrum and scale, thus providing an ideal benchmark for testing any method developed for UDG.

Metrics We plan on evaluating our proposed method on four metrics commonly used in clustering literature: clustering accuracy (ACC), normalized mutual information (NMI), Adjusted Rand Index (ARI) (Regatti et al., 2020) and Silhouette score (SIL) (Rousseeuw, 1987). ACC, NMI and ARI metrics require ground truth labels, while SIL doesn't. SIL measures how similar an image is to its own cluster compared to other clusters based on distances in the representation space. Thus, a conjunction of all 4 metrics would give us a fair estimate of the quality of the clustering achieved.

Evaluation Protocol For our experiments, we will perform leave-one-domain-out evaluation where one domain is held-out as target, commonly done in DG literature. Multiple source domains may or may not be used during training. Due to the unavailability of labeled data anywhere in the UDG setting, we cannot perform validation. We thus train all models until convergence on the source data and use the trained model as a fixed feature extractor for the target domain.

Implementation Details We used a randomly initialized ResNet-18 (He et al., 2016) model as our backbone network for all experiments. We performed a grid search for hyperparameters until we found a plateau in the training loss. We searched in the following ranges – {SGD, Adam} optimizer with learning rate $\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and mini-batch size of {128, 256}, parameters α and β to $\{1e^{-4}, 1e^{-5}\}$ and {0.1, 1.0} respectively. We also randomly crop the images, flip them horizontally and apply colour jitter. We run all our experiments across three seeds and report the mean.

4. Additional Experiments

We plan to conduct additional experiments to ascertain the robustness of the proposed algorithm as well as gain useful insight into the method. (i) To test whether the proposed curriculum strategy is useful, we will train and test on a single domain so that domain shift is controlled for. (ii) We also plan to test whether the proposed method is able to discover clusters for novel classes in the target domain that are not present in the source domain (iii) We could extend our work to settings where labeled data is available for a few domains by adding a cross-entropy loss term for those domains.

Best Practices Along with the basic meta-learning based solution that we propose, we expect to use the following best practices common in the domain shift literature. (i) **Domain-specific Batch Normalization** (Maria Carlucci et al., 2017): We will compute batch statistics separately for each domain. The motivation is that the domain-shift can be reduced by aligning the different source feature distributions to a Gaussian reference distribution. (ii) **Domain Randomization** (Volpi et al., 2021): Meta-learning approaches usually work better when there is a diverse range of tasks to learn from. In the same vein, we propose to use domain randomization (DR) to artificially increase the number of meta-source domains.

Visualizations We plan to visualize the representation space using t-SNE embeddings across training to monitor whether the model is getting progressively better as well as compare the final clustering after convergence with baseline methods. Another plot that might provide some insight into the method is to plot the meta-loss versus the clustering accuracy of the target domain per epoch similar to the one provided in (Wu et al., 2018).

5. Results

In this section, we (i) summarize our experimental results, (ii) discuss our findings in the context of our proposed hypothesis on handling domain shift and cluster imbalance, and (iii) document the modifications to the original protocol.

5.1. Main Experiments

Table 2 compares the clustering accuracy on three benchmark datasets for the baseline methods and our method. For the self-supervised learning (SSL) baseline, we used SimCLR (Chen et al., 2020). We observed that training a SimCLR model on the aggregated source domains, and then using this trained model as a feature extractor for the target domain along with running K-means on these feature representations gave the best results. This relatively simple baseline performed better than the ACIDS w/o TA baseline which is specifically designed for learning to cluster under domain shift. We experimented with two state-of-the-art deep clustering methods, one from each category – (i) Instance Discrimination & Feature Decorrelation (IDFD) (Tao et al., 2021), which learns a representation space that is amenable to being clustered using a simple technique like K-means (ii) SCAN (Van Gansbeke et al., 2020), which directly outputs a probability distribution over clusters. Both methods performed worse than SimCLR by significant margins. For SCAN, we did not perform the self-labeling step for fair comparison with our meta-learning algorithm, as it cannot be easily incorporated in an end-to-end fashion in our proposed method. We do however initialize the weights of the model with SimCLR as in the original paper. We ran our proposed meta-clustering (Algorithm 1) on top of each of the above two deep clustering methods, denoted as Meta-Clustering (IDFD) and & Meta-Clustering (SCAN) in our results. For both methods, using meta-learning did not lead to an improvement over the base clustering methods or the SimCLR baseline on average.

Our proposed curriculum learning strategy (Algorithm 2) can be tested only on deep clustering algorithms which directly output a probability distribution over clusters, therefore we test it with SCAN. The results indicate that the curriculum learning strategy is not improving over meta-clustering across the experiments too.

We train models on source domains until convergence and use that model as a fixed feature extractor for retrieving the target domain embeddings for the Photo class in the PACS dataset and visualize them using t-SNE in Figure 2. The visualization for SimCLR indicates that it is doing a better job of semantically clustering the target domain images into their respective classes, whereas no clear pattern is observed for other methods. We provide results on additional metrics in Appendix A.

5.2. Findings

Our experiments suggest that *having good, generalizable features is more important than learning to cluster or explicitly handle domain shift in the UDG setting.* Sophisticated deep clustering algorithms which were designed without keeping domain shift in mind did not fare well. In fact, algorithms such as ACIDS w/o TA and our proposed learning-to-learn method, which were specifically designed for learning to cluster under domain shift also performed worse than the relatively simpler SimCLR baseline. It is worth noting that SimCLR doesn't have any components that explicitly handle *domain shift* or *clustering.* However, its capability of learning effective representations is valuable for this task. In the light of our hypotheses, the results seem to indicate that learning a good, general-purpose feature representation is more valuable than explicitly addressing domain shift or clustering as the end objective. It is possible that however better methods to handle domain shift and clustering, perhaps aligned with the SimCLR methodology, could further improve upon the representations learned through such an approach.

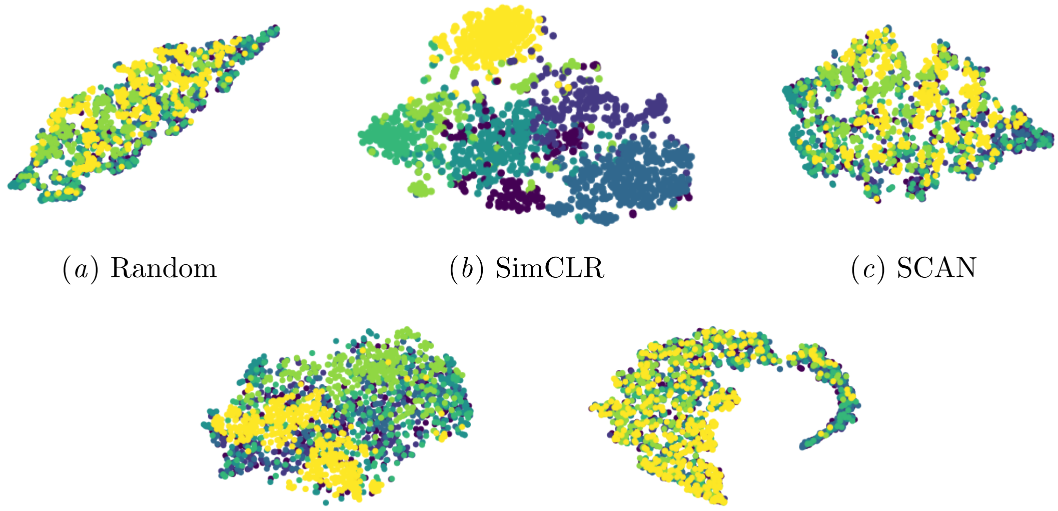
Our findings are similar to recent studies in Unsupervised Domain Adaptation (UDA) (Shen et al., 2022) and Domain Generalization (DG) (Gulrajani and Lopez-Paz, 2020). Shen et al. (2022) observe that contrastive pre-training on unlabeled source and target data

Table 2: Clustering Accuracy on PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017) & Office (Saenko et al., 2010) averaged across three runs. The best results for each dataset are highlighted in bold, the second best results are underlined.

PACS	Photo	Art	Cartoon	Sketch	Average
Uniform	14.3	14.3	14.3	14.3	14.3
Random + K-Means	26.2	21.9	24.7	30.8	25.9
SimCLR + K-Means	**62.6**	30.1	**38.1**	**43.5**	**43.6**
IDFD	27.5	22.9	25.0	36.0	27.9
Modified SCAN	45.4	29.8	27.9	29.1	33.1
ACIDS w/o TA	44.2	**34.8**	36.5	40.8	39.1
Meta-Clustering (IDFD)	25.9	21.9	28.0	31.0	26.7
Meta-Clustering (SCAN)	31.4	25.6	22.0	23.8	25.7
Curriculum + Meta-Clustering	23.8	26.7	20.8	26.9	24.5

OfficeHome	Product	Art	Clipart	Real	Average
Uniform	1.5	1.5	1.5	1.5	1.5
Random + K-Means	11.5	9.0	9.0	9.0	9.6
SimCLR + K-Means	**26.7**	**12.6**	**14.7**	**14.6**	**17.1**
IDFD	12.7	9.2	9.5	9.3	10.2
Modified SCAN	14.1	11.2	10.5	6.4	10.6
ACIDS w/o TA	9.3	8.8	7.8	7.4	8.3
Meta-Clustering (IDFD)	10.1	9.2	7.1	7.2	8.4
Meta-Clustering (SCAN)	9.1	8.1	7.3	6.7	7.8
Curriculum + Meta-Clustering	7.1	8.7	6.8	6.9	7.4

Office	Amazon	DSLR	Webcam	Average
Uniform	3.2	3.2	3.2	3.2
Random + K-Means	15.1	26.1	25.1	22.1
SimCLR + K-Means	**29.7**	36.2	**35.4**	**33.8**
IDFD	15.4	**37.6**	34.0	29.0
Modified SCAN	14.9	21.0	19.1	18.3
ACIDS w/o TA	19.3	22.5	19.0	20.2
Meta-Clustering (IDFD)	16.6	25.7	24.3	22.2
Meta-Clustering (SCAN)	18.3	18.7	18.0	18.3
Curriculum + Meta-Clustering	18.0	17.6	17.8	17.8


(*a*) Random (*b*) SimCLR (*c*) SCAN

(*d*) IDFD (*e*) Curr.+Meta-Learn.

Figure 2: t-SNE visualizations for the *Photo* domain in PACS dataset

and then fine-tuning on labeled source data, is competitive with strong UDA methods. A similar conclusion is drawn by Gulrajani and Lopez-Paz (2020) where they find that training a simple empirical risk minimization (ERM) model over the aggregated source domains is competitive with state-of-the-art in DG.

5.3. Documented Modifications

(i) We discarded the ImageNet baseline as it is the only method that has access to ImageNet weights which made it an unfair comparison; (ii) We subsumed the curriculum learning additional experiment in Table 2; (iii) Given the sub-par performance of our proposed curriculum learning method, we didn't expect to gain additional insights and hence didn't perform the novel class discovery and the labeled domain additional experiments.

6. Conclusion

In this paper, we proposed a novel and practical problem setting of Unsupervised Domain Generalization (UDG). We identified and analyzed two key challenges in this setting – unsupervised learning under domain shift and cluster imbalance. We proposed a curriculum learning based meta-learning strategy to address the challenges in UDG. Results suggest that our proposed algorithm is not the state-of-the-art for the considered UDG setting. We conclude that learning general purpose features using self-supervision is a strong baseline, even outperforming sophisticated methods explicitly designed to address domain shift and clustering. We hope that this work spurs further research in this exciting area.

References

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.

Fabio M. Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:2224–2233, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00233.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Dengxin Dai and Luc Van Gool. Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2018-November:3819–3824, 2018. doi: 10.1109/ITSC.2018.8569387.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Aniket Anand Deshmukh, Jayanth Reddy Regatti, Eren Manavoglu, and Urun Dogan. Representation learning for clustering via building consensus. *arXiv preprint arXiv:2105.01289*, 2021.

Ürün Dogan, Aniket Anand Deshmukh, Marcin Bronislaw Machura, and Christian Igel. Label-similarity curriculum learning. In *European Conference on Computer Vision*, pages 174–190. Springer, 2020.

Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain Generalization via Model-Agnostic Learning of Semantic Features. (NeurIPS), 2019. ISSN 1049-5258. URL http://arxiv.org/abs/1910.13580.

Yingjun Du, Xiantong Zhen, Ling Shao, and Cees GM Snoek. Metanorm: Learning to normalize few-shot batches across domains. In *International Conference on Learning Representations*, 2020.

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:2551–2559, 2015. ISSN 15505499. doi: 10.1109/ICCV.2015.293.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Sungwon Han, Sungwon Park, Sungkyu Park, Sundong Kim, and Meeyoung Cha. Mitigating embedding and class assignment mismatch in unsupervised image classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 768–784. Springer, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jeremy Howard. Imagenet subsets for clustering. https://github.com/fastai/imagenette, 2019.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels. *arXiv preprint arXiv:2003.08264*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Da Li, Yongxin Yang, Yi Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:5543–5551, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.591.

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*, 2018a.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain Generalization with Adversarial Feature Learning. *Proceedings of the IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, pages 5400–5409, 2018b. ISSN 10636919. doi: 10.1109/CVPR.2018.00566.

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2019b.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *2021 AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 475–485. Springer, 2020.

Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5067–5075, 2017.

Willi Menapace, Stéphane Lathuilière, and Elisa Ricci. Learning to cluster under domain shift. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 736–752. Springer, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. *30th International Conference on Machine Learning, ICML 2013*, 28(PART 1):10–18, 2013.

Chuang Niu and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *arXiv preprint arXiv:2103.09382*, 2021.

Jayanth Reddy Regatti, Aniket Anand Deshmukh, Eren Manavoglu, and Urun Dogan. Consensus clustering with unsupervised representation learning. *arXiv preprint arXiv:2010.01245*, 2020.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

Kendrick Shen, Robbie Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2204.00570*, 2022.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv preprint arXiv:2106.00131*, 2021.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285. Springer, 2020.

Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pages 1743–1751, 2019. doi: 10.1109/WACV.2019.00190.

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*, pages 5018–5027, 2017.

G. Volk, S. Müller, A. v. Bernuth, D. Hospach, and O. Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 285–292, 2019. doi: 10.1109/ITSC.2019.8917269.

Riccardo Volpi, Diane Larlus, and Grégory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4453, 2021.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. Domain-irrelevant representation learning for unsupervised domain generalization. *arXiv preprint arXiv:2107.06219*, 2021.

Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6277–6286, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

Table 3: NMI on PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017) and Office (Saenko et al., 2010). The best results for each dataset are highlighted in bold, the second best results are underlined.

PACS	Photo	Art	Cartoon	Sketch	Average
Random + K-Means	0.07	0.05	0.06	0.14	0.08
SimCLR + K-Means	**0.49**	0.13	**0.19**	**0.25**	**0.27**
IDFD	0.09	0.05	0.07	0.18	0.10
Modified SCAN	0.28	0.08	0.11	0.12	0.15
ACIDS w/o TA	0.27	**0.14**	0.17	0.23	0.20
Meta-Clustering (IDFD)	0.09	0.02	0.13	0.14	0.10
Meta-Clustering (SCAN)	0.11	0.05	0.02	0.01	0.05
Curriculum + Meta-Clustering	0.08	0.03	0.05	0.06	0.06
OfficeHome	Product	Art	Clipart	Real World	Average
Random + K-Means	0.26	0.23	0.20	0.20	0.22
SimCLR + K-Means	**0.43**	**0.29**	**0.3**	**0.26**	**0.32**
IDFD	0.26	0.23	0.21	0.21	0.23
SCAN	0.27	0.24	0.21	0.11	0.21
ACIDS w/o TA	0.19	0.22	0.13	0.14	0.17
Meta-Clustering (IDFD)	0.20	0.19	0.23	0.19	0.20
Meta-Clustering (SCAN)	0.21	0.21	0.17	0.10	0.17
Curriculum + Meta-Clustering	0.19	0.21	0.13	0.13	0.16
Office	Amazon	DLSR	Webcam	Average	
Random + K-Means	0.19	0.44	0.41	0.35	
SimCLR + K-Means	**0.33**	0.52	**0.54**	**0.46**	
IDFD	0.18	**0.55**	0.50	0.41	
Modified SCAN	0.17	0.36	0.29	0.26	
ACIDS w/o TA	0.24	0.38	0.27	0.30	
Meta-Clustering (IDFD)	0.22	0.42	0.39	0.34	
Meta-Clustering (SCAN)	0.23	0.32	0.30	0.28	
Curriculum + Meta-Clustering	0.21	0.33	0.31	0.29	

Table 4: ARI on PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017) and Office (Saenko et al., 2010). The best results for each dataset are highlighted in bold, the second best results are underlined.

PACS	Photo	Art	Cartoon	Sketch	Average
Random + K-Means	0.05	0.03	0.03	0.08	0.05
SimCLR + K-Means	**0.54**	**0.09**	**0.13**	**0.18**	**0.24**
IDFD	0.07	0.02	0.04	0.11	0.06
Modified SCAN	0.23	0.04	0.05	0.07	0.11
ACIDS w/o TA	0.26	0.10	0.11	0.17	0.16
Meta-Clustering (IDFD)	0.05	0.01	0.10	0.07	0.06
Meta-Clustering (SCAN)	0.07	0.02	0.01	0.01	0.02
Curriculum + Meta-Clustering	0.06	0.03	0.01	0.03	0.03
OfficeHome	Product	Art	Clipart	Real World	Average
Random + K-Means	0.03	0.01	0.02	0.02	0.02
SimCLR + K-Means	**0.14**	**0.03**	**0.05**	**0.05**	**0.07**
IDFD	0.03	0.01	0.02	0.02	0.02
SCAN	0.04	0.02	0.03	0.01	0.02
ACIDS w/o TA	0.04	0.01	0.02	0.02	0.02
Meta-Clustering (IDFD)	0.03	0.01	0.01	0.01	0.02
Meta-Clustering (SCAN)	0.02	0.01	0.01	0.01	0.01
Curriculum + Meta-Clustering	0.02	0.01	0.01	0.01	0.01
Office	Amazon	DSLR	Webcam	Average	
Random + K-Means	0.05	0.10	0.11	0.09	
SimCLR + K-Means	**0.15**	0.18	**0.24**	**0.19**	
IDFD	0.04	**0.21**	0.19	0.15	
SCAN	0.04	0.07	0.05	0.05	
ACIDS w/o TA	0.09	0.11	0.09	0.10	
Meta-Clustering (IDFD)	0.06	0.10	0.10	0.09	
Meta-Clustering (SCAN)	0.06	0.03	0.05	0.05	
Curriculum + Meta-Clustering	0.06	0.03	0.01	0.03	

Appendix A. Additional Metrics

Table 3 and 4 provide results on additional performance metrics. Across all datasets, we observe the same pattern as that of clustering accuracy with SimCLR giving the best results on average.