A New Benchmark: On the Utility of Synthetic Data with Blender for Bare Supervised Learning and Downstream Domain Adaptation

Hui Tang[1,2] and Kui Jia[1,*]
[1] South China University of Technology [2] DexForce Co. Ltd.
eehuitang@mail.scut.edu.cn, kuijia@scut.edu.cn

Abstract

Deep learning in computer vision has achieved great success with the price of large-scale labeled training data. However, exhaustive data annotation is impracticable for each task of all domains of interest, due to high labor costs and unguaranteed labeling accuracy. Besides, the uncontrollable data collection process produces non-IID training and test data, where undesired duplication may exist. All these nuisances may hinder the verification of typical theories and exposure to new findings. To circumvent them, an alternative is to generate synthetic data via 3D rendering with domain randomization. We in this work push forward along this line by doing profound and extensive research on bare supervised learning and downstream domain adaptation. Specifically, under the well-controlled, IID data setting enabled by 3D rendering, we systematically verify the typical, important learning insights, e.g., shortcut learning, and discover the new laws of various data regimes and network architectures in generalization. We further investigate the effect of image formation factors on generalization, e.g., object scale, material texture, illumination, camera viewpoint, and background in a 3D scene. Moreover, we use the simulation-to-reality adaptation as a downstream task for comparing the transferability between synthetic and real data when used for pre-training, which demonstrates that synthetic data pre-training is also promising to improve real test results. Lastly, to promote future research, we develop a new large-scale synthetic-to-real benchmark for image classification, termed S2RDA, which provides more significant challenges for transfer from simulation to reality.

1. Introduction

Recently, we have witnessed considerable advances in various computer vision applications [16, 29, 38]. However, such a success is vulnerable and expensive in that it has been limited to supervised learning methods with abundant labeled data. Some publicly available datasets exist certainly, which include a great mass of real-world images and acquire their labels via crowdsourcing generally. For example, ImageNet-1K [9] is of 1.28M images; MetaShift [26] has 2.56M natural images. Nevertheless, data collection and annotation for all tasks of domains of interest are impractical since many of them require exhaustive manual efforts and valuable domain expertise, e.g., self-driving and medical diagnosis. What's worse, the label given by humans has no guarantee to be correct, resulting in unpredictable label noise. Besides, the poor-controlled data collection process produces a lot of nuisances, e.g., training and test data aren't independent identically distributed (IID) and even have duplicate images. All of these shortcomings could prevent the validation of typical insights and exposure to new findings.

To remedy them, one can resort to synthetic data generation via 3D rendering [10], where an arbitrary number of images can be produced with diverse values of imaging factors randomly chosen in a reasonable range, i.e., domain randomization [48]; such a dataset creation pipeline is thus very lucrative, where data with labels come for free. For image classification, Peng et al. [32] propose the first large-scale synthetic-to-real benchmark for visual domain adaptation [30], VisDA-2017; it includes 152K synthetic images and 55K natural ones. Ros et al. [37] produce 9K synthetic cityscape images for cross-domain semantic segmentation. Hinterstoisser et al. [19] densely render a set of 64 retail objects for retail detection. All these datasets are customized for specific tasks in cross-domain transfer. In this work, we push forward along this line extensively and profoundly.

The deep models tend to find simple, unintended solutions and learn shortcut features less related to the semantics of particular object classes, due to systematic biases, as revealed in [14]. For example, a model basing its prediction on context would misclassify an airplane floating on water as a boat. The seminal work [14] emphasizes that shortcut opportunities are present in most data and rarely disappear by simply adding more data. Modifying the training data to block specific shortcuts may be a promising solution, e.g., making image variation factors consistently distributed

*Corresponding author.

across all categories. To empirically verify the insight, we propose to compare the traditional fixed-dataset periodic training strategy with a new strategy of training with unduplicated examples generated by 3D rendering, under the well-controlled, IID data setting. We run experiments on three representative network architectures of ResNet [18], ViT [12], and MLP-Mixer [49], which consistently show obvious advantages of the data-unrepeatable training (cf. Sec. 4.1). This also naturally validates the typical arguments of probably approximately correct (PAC) generalization [41] and variance-bias trade-off [11]. Thanks to the ideal IID data condition enabled by the well-controlled 3D rendering, we can also discover more reliable laws of various data regimes and network architectures in generalization. Some interesting observations are as follows.

- **Do not learn shortcuts!** The test results on synthetic data without background are good enough to show that the synthetically trained models do not learn shortcut solutions relying on context clues [14].
- **A zero-sum game.** For the data-unrepeatable training, IID and OOD (Out-of-Distribution [34, 55]) generalizations are some type of zero-sum game w.r.t. the strength of data augmentation.
- **Data augmentations do not help ViT much!** In IID tests, ViT performs surprisingly poorly whatever the data augmentation is and even the triple number of training epochs does not improve much.
- **There is always a bottleneck from synthetic data to OOD/real data**. Here, increasing data size and model capacity brings no more benefits, and domain adaptation [56] to bridge the distribution gap is indispensable except for evolving the image generation pipeline to synthesize more realistic images.

Furthermore, we comprehensively assess image variation factors, e.g., object scale, material texture, illumination, camera viewpoint, and background in a 3D scene. We then find that to generalize well, deep neural networks must *learn to ignore non-semantic variability*, which may appear in the test. To this end, sufficient images with different values of one imaging factor should be generated to learn a robust, unbiased model, proving the necessity of sample diversity for generalization [42, 48, 55]. We also observe that *different factors and even their different values have uneven importance to IID generalization*, implying that the underexplored weighted rendering [3] is worth studying.

Bare supervised learning on synthetic data results in poor performance in OOD/real tests, and pre-training and then domain adaptation can improve. Domain adaptation (DA) [56] is a hot research area, which aims to make predictions for unlabeled instances in the target domain by transferring knowledge from the labeled source domain. To our knowledge, there is little research on pre-training for DA [24] (with real data). We thus use the popular simulation-to-real classification adaptation [32] as a downstream task, study the transferability of synthetic data pre-trained models by comparing with those pre-trained on real data like ImageNet and MetaShift. We report results for several representative DA methods [7, 13, 40, 45, 47] on the commonly used backbone, and our experiments yield some surprising findings.

- **The importance of pre-training for DA.** DA fails without pre-training (cf. Sec. 4.3.1).
- **Effects of different pre-training schemes.** Different DA methods exhibit different relative advantages under different pre-training data. The reliability of existing DA method evaluation criteria is unguaranteed.
- **Synthetic data pre-training vs. real data pre-training.** Synthetic data pre-training is better than pre-training on real data in our study.
- **Implications for pre-training data setting.** Big Synthesis Small Real is worth researching. Pre-train with target classes first under limited computing resources.
- **The improved generalization of DA models.** Real data pre-training with extra non-target classes, fine-grained target subclasses, or our synthesized data added for target classes helps DA.

Last but not least, we introduce a new, large-scale synthetic-to-real benchmark for classification adaptation (S2RDA), which has two challenging tasks S2RDA-49 and S2RDA-MS-39. S2RDA contains more categories, more realistically synthesized source domain data coming for free, and more complicated target domain data collected from diverse real-world sources, setting a more practical and challenging benchmark for future DA research.

2. Related Works

Real Datasets. A lot of large-scale real datasets [9, 15, 25–27, 35, 43, 44] have harnessed and organized the explosive image data from the Internet or real world for deep learning of meaningful visual representations. For example, ImageNet [9] is a large-scale image database consisting of 1.28M images from 1K common object categories, and serves as the primary dataset for pre-training deep models for vision tasks. Barbu et al. [2] collect a large real-world test set for more realistic object recognition, ObjectNet, which has 50K images and is bias-controlled. MetaShift [26] of 2.56M natural images ($\sim$ 400 classes) is formed by context guided clustering of the images from GQA [22].
Synthetic Datasets. Thanks to 3D rendering [10] and domain randomization [48], synthetic data with increased

sample diversity can be generated for free now, facilitating various vision tasks. VisDA-2017 [32] is a large-scale benchmark dataset for cross-domain object classification, focusing on the simulation-to-reality shift, with 152K synthetic images and 55K natural ones across 12 categories. For cross-domain semantic segmentation, Ros et al. [37] produce 9K synthetic images by rendering a virtual city using the Unity engine; many other works [36, 52–54] focus on computer graphics simulations to synthesize outdoor or indoor scene images. For retail detection, Hinterstoisser et al. [19] densely render 64 retail objects and a large dataset of 3D background models. For car detection, domain randomization is also utilized and developed in [33, 50].

Domain Adaptation. Domain adaptation is a developing field with a huge diversity of approaches. A popular strategy is to explicitly model and minimize the distribution shift between the source and target domains [7, 13, 31, 40, 46, 57], such that the domain-invariant features can be learned. Differently, works of another emerging strategy [8, 28, 45, 47] take steps towards implicit domain adaptation, without explicit feature alignment. In this work, we consider these representative DA methods for the empirical study, and broader introductions to the rich literature are provided in [39, 56].

3. Data Synthesis via Domain Randomization

We adopt the widely used 3D rendering [10] in a simulator for data synthesis and generate synthetic RGB images for model training. To increase sample diversity for better generalization, we apply domain randomization [48] during rendering, whose efficacy has been demonstrated in various applications [32, 33, 50]. Specifically, we start by sampling a 3D object model from one specific class of interest from ShapeNet repository [5] and place it in a blank scene; next, we set the lighting condition with a point source of randomized parameters and place the camera at random positions on the surface of a sphere of random radius, which has lens looking at the object and the intrinsic resolution of 256×256; next, we apply random materials and textures to the object; then, we use an RGB renderer to take pictures from different camera viewpoints in the configured scene; finally, the rendered images are composed over a background image chosen at random from open resources. The synthesized images with the automatically generated ground truth class labels are used as low-cost training data.

The changing ranges of image variation factors in the 3D scene are as follows. **(1)** The scale of the object in an image depends on its distance to the camera, namely the radius of a sphere on whose surface the camera is located. The radius is ranged in $[0.8, 2.4]$. **(2)** The material texture of an object is from CCTextures [20], which contains actual images. **(3)** The point light is on a shell centered at $[1, 2, 3]$, whose radius and elevation are ranged in $[1, 7]$ and $[15, 70]$ respectively. **(4)** The camera lies on the surface of a sphere centered at $[0, 0, 0]$ (also the object center) and its azimuth and elevation are from $0°$ to $360°$. **(5)** The background images are from Haven [21], which includes environment HDRs.

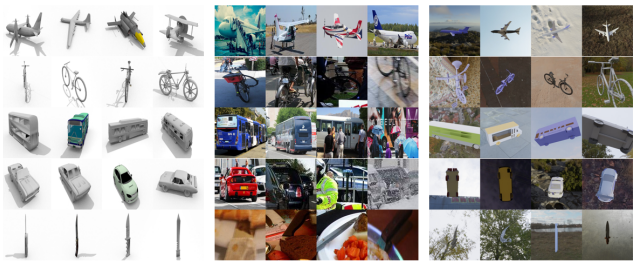


Figure 1. Sample images from the training (left) and validation (middle) domains of VisDA-2017 and our synthesized data (right).

Remarks. It is noteworthy that VisDA-2017 [32] generates synthetic images by rendering 3D models just under varied camera angles and lighting conditions. Differently, we vary the values of much more image variation factors, leading to more realistic and diverse samples, as illustrated in Fig. 1.

4. Experiment and Evaluation

We empirically demonstrate the utility of our synthetic data for supervised learning and downstream transferring by exploring the answers to the following questions:

- Can we utilize synthetic data to verify typical theories and expose new findings? What will we find when investigating the learning characteristics and properties of our synthesized new dataset comprehensively?
- Can a model trained on non-repetitive samples converge? If it could, how will the new training strategy perform when compared to fixed-dataset periodic training? Can the comparison provide any significant intuitions for shortcut learning and other insights?
- How will the image variation factors in domain randomization affect the model generalization? What new insights can the study provide for 3D rendering?
- Can synthetic data pre-training be on par with real data pre-training when applied to downstream synthetic-to-real classification adaptation? How about large-scale synthetic pre-training with a small amount of real data?
- Is our S2RDA benchmark more challenging and realistic? How does it differ from VisDA-2017?

4.1. Empirical Study on Supervised Learning

Data Settings. We use the 10 object classes common in ShapeNet and VisDA-2017 for the empirical study. We term the synthetic and real domains of the 10 classes in VisDA-2017 as SubVisDA-10. For the traditional fixed-dataset periodic training, we generate 12K synthetic images in each class and train the model on the dataset with fixed size epoch by epoch. For our used sample-unrepeatable training,

we have mutually exclusive batches of synthesized samples per iteration. At inference, we evaluate the learned model on three types of test data: IID data of 6K samples per class which follow the same distribution as the synthesized training data, IID data of 60K images without background to examine the dependency of network predictions on contexts, and OOD data, i.e., real images from SubVisDA-10. For training, we consider three data augmentation strategies with different strengths: no augmentation which has only the center crop operation, weak augmentation based on pixel positions such as random crop and flip [9], and strong augmentation which transforms both position and value of pixels in an image, e.g., random resized crop and color jitter [6]. In the test phase, we use no augmentation.

Implemental Details. We adopt the standard cross-entropy loss function for pattern learning. We do experiments using ResNet-50 [18], ViT-B [12], and Mixer-B [49] as the backbone. We train the model from scratch for 200K iterations and use the SGD optimizer with batch size 64 and the cosine learning rate schedule to update network parameters. We report the overall accuracy (Acc. %) or mean class precision (Mean %) at the same fixed random seed across all experiments. Other settings are detailed in the appendix.

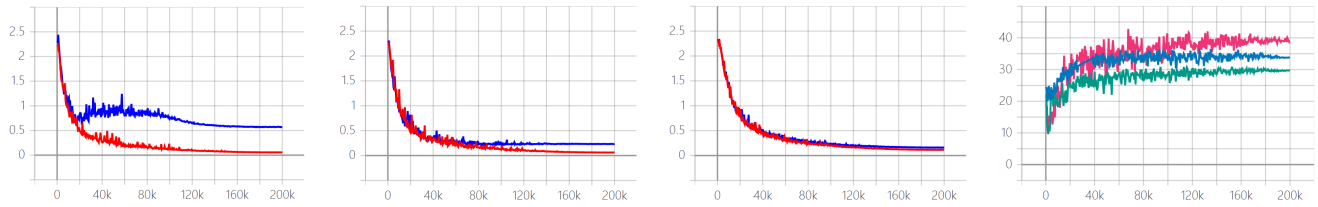
Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
Backbone: ResNet-50 (23.53M)						
SubVisDA-10	T	N	11.25	11.72	22.02	14.71
Ours	T	N	87.63	78.55	23.35	23.36
Ours	F	N	**98.19**	**96.39**	**25.04**	**26.05**
SubVisDA-10	T	W	12.31	13.53	25.95	16.83
Ours	T	W	95.54	91.37	23.97	22.89
Ours	F	W	**98.10**	**96.35**	**27.47**	**27.49**
SubVisDA-10	T	S	17.39	20.32	33.07	27.48
Ours	T	S	94.86	95.33	42.24	41.73
Ours	F	S	**96.26**	**96.50**	**42.82**	**42.25**
Backbone: ViT-B (85.78M) †: Training for 600K iterations						
SubVisDA-10	T	N	12.68	11.30	24.28	17.81
Ours	T	N	68.51	61.50	26.65	24.13
Ours†	T	N	70.58	62.15	26.57	24.23
Ours	F	N	**76.34**	**71.46**	**30.10**	**26.93**
SubVisDA-10	T	W	11.77	11.20	26.53	19.22
Ours	T	W	72.79	67.46	**30.04**	26.45
Ours	F	W	**73.93**	**68.59**	29.92	**26.80**
SubVisDA-10	T	S	14.45	12.89	31.52	23.74
Ours	T	S	62.85	63.96	**31.79**	**26.56**
Ours	F	S	**64.26**	**64.30**	30.89	26.28
Backbone: Mixer-B (59.12M)						
SubVisDA-10	T	N	12.85	15.17	21.56	17.02
Ours	T	N	66.05	57.66	21.85	21.22
Ours	F	N	**90.22**	**85.86**	**28.54**	**27.98**
SubVisDA-10	T	W	13.99	23.12	27.67	19.86
Ours	T	W	78.43	71.48	27.15	26.01
Ours	F	W	**90.32**	**86.13**	**29.11**	**29.49**
SubVisDA-10	T	S	14.88	24.85	33.19	26.12
Ours	T	S	81.72	83.06	**36.57**	33.43
Ours	F	S	**84.16**	**85.25**	36.50	**33.75**

Table 1. Training on a fixed dataset vs. non-repetitive samples. FD: Fixed Dataset, True (T) or False (F). DA: Data Augmentation, None (N), Weak (W), or Strong (S). BG: BackGround.

4.1.1 Results

Fixed-Dataset Periodic Training vs. Training on Non-Repetitive Samples. Under the ideal IID data condition enabled by 3D rendering, we empirically verify significant insights on shortcut learning [14], PAC generalization [41], and variance-bias trade-off [11], by making comparisons between fixed-dataset periodic training and training on non-repetitive samples. Results are shown in Table 1 and Figs. 2 and A1 (learning process in the appendix). We highlight several observations below. **(1)** With more training data of increased sample diversity, the data-unrepeatable training exhibits *higher generalization accuracy and better convergence performance* than the fixed-dataset training. **(2)** To intuitively understand what the models have learned, we visualize the saliency/attention maps in Figs. A2-A5. We observe that all models attend to image regions from global (context) to local (object) as the learning process proceeds; the data-unrepeatable training achieves *qualitative improvements* over the fixed-dataset training. **(3)** Our synthesized data used for training yield *higher OOD test accuracy* than SubVisDA-10 as they share more similarities to the real data, as shown in Fig. 1. **(4)** The fixed-dataset training displays *overfitting phenomenons* whilst the data-unrepeatable training *does not* (cf. (a-d) in Fig. A1), since the former samples training instances from an empirical distribution with high bias and low variance, and thus cannot perfectly generalize to the unseen test instances sampled from the true distribution. **(5)** With strong data augmentation, the data-unrepeatable training has the test results on IID w/o BG data not only at their best but also better than those on IID data, implying that the trained models *do not learn shortcut solutions that rely on context clues in the background*.

Evaluating Various Network Architectures. In addition to Table 1, we also show the learning process of various network architectures in Figs. 2d and A6. We take the following interesting observations. **(1)** On the fixed-dataset training and IID tests, ViT-B *performs surprisingly poorly* whatever the data augmentation is, when compared with ResNet-50; even the triple number of training epochs does not work as well as expected (e.g., in [12]). **(2)** When training on non-repetitive images without strong data augmentation, ViT-B and Mixer-B perform better than ResNet-50 in OOD tests whereas they perform much worse with strong data augmentation. Maybe they are more suitable for handling data with a certain (or smaller) range of diversity. Namely, *different network architectures have different advantages for different data augmentations*, suggesting that neural architecture search (NAS) should also consider the search for data augmentation. **(3)** With strong data augmentation, ResNet-50 fits best and shows the best convergence, though it has a more volatile learning process for the OOD test (cf. Figs. 2d and A6). **(4)** ResNet-50 produces more accurate saliency map visualizations, where the attended regions are semantically related (cf. Figs. A2-A5).



(a) Test loss (IID) (b) Test loss (IID) (c) Test loss (IID) (d) Test acc. (OOD)

Figure 2. Learning process. **(a-c):** Training ResNet-50 on a fixed dataset (blue) or non-repetitive samples (red) for no, weak, and strong data augmentations. **(d):** Training ResNet-50 (red), ViT-B (green), and Mixer-B (blue) on non-repetitive samples with strong data augmentation.

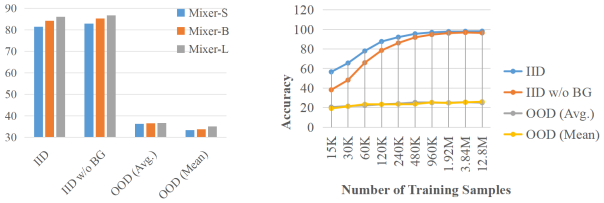


Figure 3. Generalization accuracy w.r.t. model capacity (**a/left**) or training data quantity (**b/right**).

Impact of Model Capacity. When Mixer is used as the backbone for data-unrepeatable training with strong data augmentation, we do experiments by varying the number of layers in $[8, 12, 24]$, i.e., Mixer-S (18.02M), Mixer-B (59.12M), and Mixer-L (207.18M). The results are shown in Fig. 3a. Although better results are achieved by higher capacity models, the performance gain is *less and less significant*. It also suggests that given a specific task, one model always has a bottleneck, which may be broken by adjusting the training data and learning algorithm.

Impact of Training Data Quantity. We do experiments by doubling the number of training samples and the last 12.8M is the upper bound. We use the fixed-dataset periodic training with ResNet-50 and no data augmentation, and show the results in Fig. 3b. **(1)** In IID tests, as the number of training samples is continuously doubled, the performance gets better and is near perfect finally. **(2)** In real OOD tests, the performance gain is slighter and along the last 4 numbers, the performance gain almost disappears under our data regime. It demonstrates that simply generating more synthesized images may *get stuck* at last and one can resort to domain adaptation approaches to reduce the distribution shift or more realistic simulation for image synthesis.

Impact of Data Augmentations. Data augmentation plays an important role in deep learning and we separately analyze its impact. There are a few noteworthy findings in Table 1. **(1)** For training ResNet-50 on a fixed dataset, weak augmentation can enable the learnability from our synthesized data to IID data (e.g., $> 95\%$); from synthetic to IID w/o BG, strong augmentation is necessary; from synthetic to OOD, strong augmentation has the highest learnability. These observations enlighten us: *given a limited set of training data, is there necessarily some kind of data augmentation that makes it learnable from training to test?* **(2)** For the data-unrepeatable training, the results in IID tests *get worse* while those in OOD tests *get better* when strengthening the data augmentation, since the distribution of strongly augmented training data *differs* from that of IID test data but is *more similar* to that of OOD test data.

Object Scale			**Material Texture**		
Value	IID	IID w/o BG	Value	IID	IID w/o BG
1	68.77	58.00	Metal	79.58	68.78
1.5	80.80	72.22	Plastic	50.29	46.82
2	77.61	70.10	Fingerprints	50.35	62.27
Mix	87.12	77.55	Moss	68.62	63.93
Illumination			**Camera Viewpoint**		
Value	IID	IID w/o BG	Value	IID	IID w/o BG
Location 1	86.48	76.02	Location 1	24.60	26.56
Location 2	86.60	76.75	Location 2	27.21	28.88
Radius	86.91	78.83	Location 3	32.82	32.76
Elevation	87.12	77.39	Location 4	33.79	33.07
Background			**Full Randomization**		
Value	IID	IID w/o BG	Value	IID	IID w/o BG
No Background	17.68	**94.75**	Random	**87.63**	78.55

Table 2. Fix vs. randomize image variation factors (ResNet-50).

4.2. Assessing Image Variation Factors

To know *how individual image variation factors in domain randomization affect the model generalization*, we do an ablation study by fixing one of them at a time. For each variant, we accordingly synthesize 120K training images and train a ResNet-50 from scratch with no data augmentation. Other implementation details are the same as those in Sec. 4.1. We compare the one-fixed variants with the baseline of full randomization and report results in Table 2.

Object Scale. The scale of target object in an image is changed by modifying the distance between the object and camera in the virtual 3D scene, whose value is set as 1, 1.5, 2, or a mix of the three. We can observe that when the object scale is fixed to one value, the recognition accuracy in IID and IID w/o BG tests drops significantly, e.g., by 18.86% and 20.55% respectively; setting the object-to-camera distance as 1.5 achieves the best performance among the three scales; mixing the three scales restores most of the baseline performance. The observations show that *decreasing the sample diversity would damage the generalization per-*

Pre-training Data	# Iters	# Epochs	No Adaptation		DANN		MCD		RCA		SRDC		DisClusterDA	
			Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean
No Pre-training	-	-	23.89	14.21	22.30	17.72	17.99	16.20	19.15	15.19	19.58	15.92	20.87	17.31
Ours	200K	107	47.73	42.96	47.91	48.94	55.23	56.86	54.27	52.72	44.70	47.45	54.09	54.91
Ours w. SelfSup	200K	107	47.80	42.81	47.25	48.32	56.71	58.33	53.44	53.31	40.21	40.50	54.37	54.15
SynSL	200K	1	47.50	44.12	47.41	49.48	55.06	56.92	53.61	53.50	36.30	37.57	53.10	54.88
SynSL	1.2M	6	51.22	48.57	55.90	56.50	64.52	67.70	58.19	59.67	51.87	52.32	61.32	63.70
SynSL	2.4M	12	53.47	**53.84**	59.59	59.11	65.55	68.83	60.47	61.19	55.17	58.10	63.62	64.89
SynSL	4.8M	24	55.02	53.72	60.55	**60.78**	**65.69**	**69.53**	60.33	60.05	55.80	58.36	64.01	**65.36**
SubImageNet	200K	499	42.74	37.16	49.64	45.43	54.88	52.12	58.24	55.45	56.78	51.24	56.21	51.09
Ours+SubImageNet	200K	88	49.61	47.88	55.35	56.22	60.90	62.16	61.11	60.31	60.07	61.74	62.22	62.47
ImageNet-990	200K	10	31.91	26.31	34.68	32.29	39.48	37.84	45.10	43.10	43.69	40.95	41.56	39.40
ImageNet-990+Ours	200K	9	36.53	30.58	38.15	35.22	42.38	41.84	46.19	43.45	45.87	42.95	42.07	39.40
ImageNet	200K	10	40.37	33.25	42.57	40.22	49.04	47.86	52.36	47.90	51.62	47.88	49.29	46.17
ImageNet	1.2M	60	54.69	51.27	58.50	56.02	65.28	65.88	62.69	60.28	60.33	55.00	62.28	61.00
ImageNet	2.4M	120	53.84	47.55	58.45	55.38	65.27	65.38	61.65	60.82	61.65	56.30	62.02	60.46
ImageNet★	600K	120	**57.10**	51.83	**61.92**	58.75	64.59	64.87	**67.72**	**66.17**	**69.00**	**64.92**	**68.35**	64.65
MetaShift	200K	5	38.18	30.31	38.29	34.04	45.39	43.63	45.93	42.67	42.83	38.02	40.17	35.09
MetaShift	1.2M	30	48.00	39.99	53.00	48.17	64.04	61.30	53.97	51.09	48.69	44.49	60.28	57.15
MetaShift	2.4M	60	47.24	39.21	58.41	53.85	61.10	58.52	58.64	55.35	51.71	47.29	62.71	60.18

Table 3. Domain adaptation performance on SubVisDA-10 with varied pre-training schemes (ResNet-50). ★: Official checkpoint. Green or red: Best Acc. or Mean in each row (among compared DA methods). Ours w. SelfSup: Supervised pre-training + contrastive learning [6].

formance; different scales have different importance.

Material Texture. We fix the material texture of target object in an image as metal, plastic, fingerprints, or moss. We observe that compared to full randomization, fixing the material texture degenerates the performance largely in IID and IID w/o BG tests, e.g., by 37.34% and 31.73% respectively.

Illumination. We change the illumination by fixing the light location to [5, −5, 5] or [4, 5, 6] or narrowing the range of radius or elevation to [3, 4] and [20, 30]. We find that the location of light source has a greater influence than its radius and elevation. Compared with other factors, illumination has much less impact on class discrimination.

Camera Viewpoint. Recall that the camera always looks at the target object. We change the viewpoint of camera by placing it in different 3D locations: [0, 1, 1], [0, -1, 1], [1, 0, 1], or [-1, 0, 1]. We can see that fixing the camera viewpoint makes the results in IID and IID w/o BG tests deteriorate greatly, e.g., by 63.03% and 51.99% respectively.

Background. When the background is lacking in an image, the accuracy in the IID test suffers an abrupt decrease of 69.95% while that in the IID w/o BG test improves by 16.2% due to reduced distribution shift. Among 5 factors, the background is *the most important* for IID generalization.

Remarks. Different rendering variation factors and even their different values have uneven importance to model generalization. It suggests that the under-explored direction of weighted rendering [3] is worth studying, and our results in Table 2 provide preliminary guidance/prior knowledge for learning the distributions of variation factors.

4.3. Exploring Pre-training for Domain Adaptation

Data Settings. The data used for pre-training can be selected in several datasets and their variants: **(1)** our synthesized 120K images of the 10 object classes shared by SubVisDA-10 (Ours), **(2)** our synthesized 12.8M images of the 10 classes (for supervised learning, termed SynSL), **(3)** the subset collecting examples of the 10 classes from ImageNet [9] (25,686 images, termed SubImageNet), **(4)** our synthesized 120K images combined with SubImageNet, **(5)** ImageNet-990, where the fine-grained subclasses for each of the 10 classes are merged into one, **(6)** ImageNet-990 combined with our 120K synthetic images, **(7)** the full set of ImageNet (1K classes), and **(8)** MetaShift [26] (2.56M). For fine-tuning, we use domain adaptation (DA) on SubVisDA-10 as the downstream task, which comprises 130,725 labeled instances in the source domain and 46,697 unlabeled ones in the target domain. We follow the standard DA training protocol [13, 40]. We report classification results of overall accuracy (Acc. %) and mean class precision (Mean %) on the target domain under a fixed random seed.

Implemental Details. For domain adaptation, we use DANN [13], MCD [40], RCA [7], SRDC [45], and DisClusterDA [47] as baselines. We closely follow the specific algorithms in the respective papers of these baselines. We use a pre-trained ResNet-50 as the base model. We train the model for 20 epochs with batch size 64 via SGD. Refer to Sec. 4.1 and the appendix for other details.

4.3.1 Results and Discussions

To *study the effects of pre-training on synthetic-to-real adaptation*, we examine several DA methods when varying the pre-training scheme in terms of pre-training data and duration. The results are reported in Table 3 and Figs. 4 and A7. We emphasize several remarkable findings below.

The importance of pre-training for DA. DA fails without pre-training. With no pre-training, the very baseline No Adaptation that trains the model only on the labeled source data, outperforms all compared DA methods in overall accuracy, despite the worst mean class precision. It verifies that pre-training is indispensable for DA and involving the

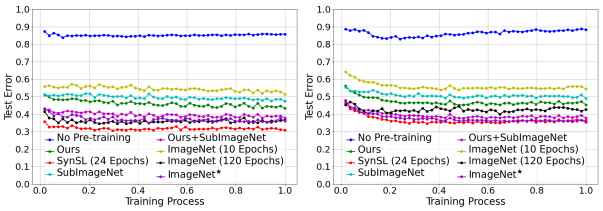


Figure 4. Learning process (Mean) of MCD (**left**) and DisClusterDA (**right**) when varying the pre-training scheme.

target data in training may alleviate class imbalance.

Effects of different pre-training schemes. Different DA methods exhibit different relative advantages under different pre-training data. When pre-training on our synthesized data, MCD achieves the best results; when pre-training on Ours+SubImageNet, DisClusterDA outperforms the others; when pre-training on ImageNet★, SRDC yields the best performance. What's worse, the reliability of existing DA method evaluation criteria is unguaranteed. With different pre-training schemes, the best performance is achieved by different DA methods. When pre-training on ImageNet for 10, 60, or 120 epochs, the best results are achieved by RCA, MCD, and SRDC respectively; when pre-training on MetaShift for 5, 30, or 60 epochs, the best results are achieved by RCA, MCD, and DisClusterDA respectively.

Synthetic data pre-training vs. real data pre-training. Synthetic data pre-training is better than pre-training on real data in our study. With the same 200K pre-training iterations, our synthetic data often bring more benefits than real data from ImageNet or MetaShift, though the top-ranked performance is achieved by extending the pre-training time on real data. Under the same experimental configuration, SynSL pre-training for 24 epochs is comparable to or better than pre-training on ImageNet for 120 epochs and maybe it's because SynSL is 10 times ImageNet's size. The observation indicates that with our 12.8M synthetic data pre-training, the DA methods can yield the new state of the art.

Implications for pre-training data setting. Big Synthesis Small Real is worth deeply researching. Ours+SubImageNet augmenting our synthetic data with a small amount of real data, achieves remarkable performance gain over Ours, suggesting a promising paradigm of supervised pre-training — Big Synthesis Small Real. On the other hand, pre-train with target classes first under limited computing resources. With 200K pre-training iterations, SubImageNet performs much better than ImageNet (10 Epochs), suggesting that one should consider pre-training with target classes first in cases of low computation budget, e.g., real-time deployment on low-power devices like mobile phones, laptops, and smartwatches. Here, we have two questions to be considered: do we have unlimited computing resources for pre-training? Is the domain-specific pre-training more suitable for some industrial applications?

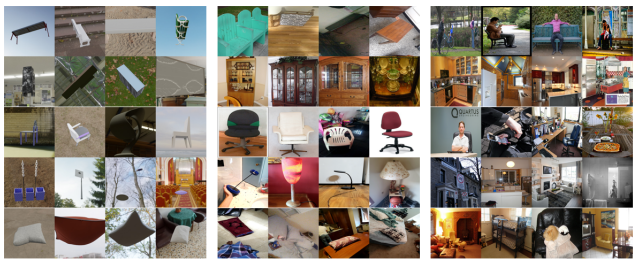


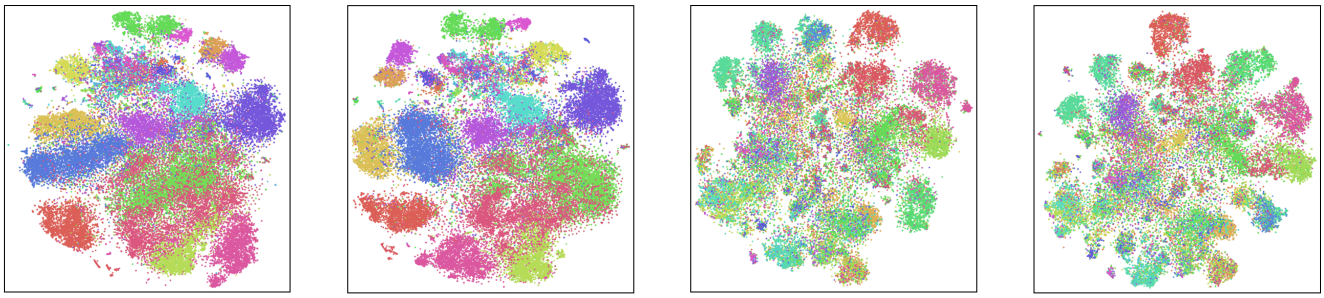
Figure 5. Sample images from the synthetic (left) domain and the real domains of S2RDA-49 (middle) and S2RDA-MS-39 (right).

The improved generalization of DA models. Real data pre-training with extra non-target classes, fine-grained target subclasses, or our synthesized data added for target classes helps DA. ImageNet (120 Epochs) involving both target and non-target classes in pre-training is better than SubImageNet involving only target classes, indicating that learning rich category relationships is helpful for downstream transferring. with 200K pre-training iterations, ImageNet-990 performs much worse than ImageNet, implying that pre-training in a fine-grained visual categorization manner may bring surprising benefits. Ours+SubImageNet adding our synthesized data for target classes in SubImageNet, produces significant improvements and is close to ImageNet (120 Epochs); ImageNet-990+Ours improves over ImageNet-990, suggesting that synthetic data may help improve the performance further.

Convergence analysis. In Figs. 4 and A7, the convergence from different pre-training schemes for the same DA method differs in speed, stability, and accuracy. In Fig. 4, SynSL with 24 epochs outperforms ImageNet with 120 epochs significantly; notably, SynSL is on par with or better than ImageNet★, supporting our aforementioned findings.

4.4. A New Synthetic-to-Real Benchmark

Dataset Construction. Our proposed Synthetic-to-Real benchmark for more practical visual DA (termed S2RDA) includes two challenging transfer tasks of S2RDA-49 and S2RDA-MS-39 (cf. Fig. 5). In each task, source/synthetic domain samples are synthesized by rendering 3D models from ShapeNet [5] (cf. Sec. 3). The used 3D models are in the same label space as the target/real domain and each class has 12K rendered RGB images. The real domain of S2RDA-49 comprises 60,535 images of 49 classes, collected from ImageNet validation set, ObjectNet [2], VisDA-2017 validation set, and the web [1]. For S2RDA-MS-39, the real domain collects 41,735 natural images exclusive for 39 classes from MetaShift [26], which contain complex and distinct contexts, e.g., object presence (co-occurrence of different objects), general contexts (indoor or outdoor), and object attributes (color or shape), leading to a much harder task. In Fig. A8, we show the long-tailed distribution



(a) No Adaptation (b) SRDC (c) No Adaptation (d) DisClusterDA

Figure 6. The t-SNE visualization of target domain features extracted by different models on S2RDA-49 (a-b) and S2RDA-MS-39 (c-d).

Transfer Task	No Adaptation		DANN		MCD		RCA		SRDC		DisClusterDA	
	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean
S2RDA-49	51.89	42.19	47.06	47.64	42.51	47.77	47.07	48.46	**61.52**	**52.98**	53.03	52.34
S2RDA-MS-39	22.03	20.54	22.82	22.20	22.07	22.16	23.34	22.53	25.83	24.55	**27.14**	**25.33**

Table 4. Domain adaptation performance on S2RDA (ResNet-50).

of image number per class in each real domain. Compared to VisDA-2017 [32], our S2RDA contains more categories, more realistically synthesized source domain data coming for free, and more complicated target domain data collected from diverse real-world sources, setting a more practical, challenging benchmark for future DA research.

Benchmarking DA Methods. We use ResNet-50 as the backbone, initialized by the official ImageNet pre-trained checkpoint [18]. Refer to Sec. 4.3 and the appendix for other implementation details. We report the results on S2RDA in Table 4 and show t-SNE [51] visualizations in Fig. 6. In the overall accuracy (Acc.) on S2RDA-49, the adversarial training based methods DANN, MCD, and RCA perform worse than No Adaptation, demonstrating that explicit domain adaptation could deteriorate the intrinsic discriminative structures of target domain data [45, 47], especially in cases of more categories; in contrast, SRDC produces significant quantitative and qualitative improvements, 14.45% and 8.49% higher than RCA and DisClusterDA respectively, but 7.48% lower than Acc. on SubVisDA-10 (cf. Table 3), indicating the difficulty of the S2RDA-49 task. On S2RDA-MS-39, the classification accuracy is much worse than that on S2RDA-49 (decrease of more than 20%), and the compared methods show much less performance difference; the highest accuracy is only 27.14% achieved by DisClusterDA, showing that S2RDA-MS-39 is a very challenging task. To summarize, *domain adaptation is far from being solved* and we expect that our results contribute to the DA community as new benchmarks, though more careful studies in different algorithmic frameworks are certainly necessary to be conducted.

5. Conclusions and Future Perspectives

This paper primarily aims to publish new datasets including our synthetic dataset SynSL (12.8M) and S2RDA, and benchmarks the datasets via supervised learning and downstream transferring. In the context of image classification, our work is the first comprehensive study on synthetic data learning, which is completely missing. We propose exploiting synthetic datasets to explore questions on model generalization and benchmark pre-training strategies for DA. We build randomized synthetic datasets using a 3D rendering engine and use this established system to manipulate the generation of images by altering several imaging factors. We find that synthetic data pre-training has the potential to be better than pre-training on real data, our new benchmark S2RDA is much more practical for synthetic-to-real DA, to name a few. We expect that these results contribute to the transfer learning community as new benchmarks, though the research on more synthetic datasets, more models, and more DA methods is certainly to be done.

Synthetic data as a new benchmark. Synthetic data are well suited for use as toy examples to verify existing deep learning theoretical results or explore new theories.

Evaluation metrics robust to pre-training. The comparison among various DA methods yields different or even opposite results when using different pre-training schemes (cf. Sec. 4.3). DA researchers should propose and follow evaluation metrics enabling effective and fair comparison.

More realistic simulation synthesis. We will consider more imaging parameters, e.g., randomizing the type and hue of the light, including 77 physical objects with actual textures from YCB [4], and using the flying distractor [23].

To explore deep learning based data generation. Our proposed paradigm of empirical study can generalize to any data generation pipeline. Our findings may be data source specific and the generalizability to other pipelines like GANs, NeRFs, and AutoSimulate [3] is to be explored.

Applicability to other vision tasks. Our new paradigm of empirical study for image classification can also be applied to other vision tasks of semantic analysis, e.g., Kubric [17] and HyperSim [36] for segmentation and object detection.

Acknowledgments. This work is supported in part by Program for Guangdong Introducing Innovative and Entrepreneurial Teams (No.: 2017ZT07X183) and Guangdong R&D key project of China (No.: 2019B010155001).

References

[1] NIKHIL AKKI. Furniture detector. In *CC0: Public Domain*. 7

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Proc. Neur. Info. Proc. Sys.*, volume 32, 2019. 2, 7

[3] Harkirat Singh Behl, Atilim Güneş Baydin, Ran Gal, Philip H. S. Torr, and Vibhav Vineet. Autosimulate: (quickly) learning synthetic data generation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proc. Eur. Conf. Comput. Vis.*, pages 255–271, Cham, 2020. Springer International Publishing. 2, 6, 8

[4] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, , and Aaron M. Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. In *IEEE Robotics and Automation Magazine*, pages 36–52, 2015. 8

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3, 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn.*, 2020. 4, 6

[7] S. Cicek and S. Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1416–1425, 2019. 2, 3, 6

[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3941–3950, 2020. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, 2009. 1, 2, 4, 6

[10] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019. 1, 2, 3

[11] Pedro Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55:78–87, 2012. 2, 4

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. on Learn. Rep.*, 2021. 2, 4

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journ. of Mach. Learn. Res.*, 17:2096–2030, 2016. 2, 3, 6

[14] R. Geirhos, JH. Jacobsen, C. Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2:665–673, 2020. 1, 2, 4

[15] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 2

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 580–587, 2014. 1

[17] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3761, June 2022. 8

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016. 2, 4, 8

[19] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Marek Martina, and Martin Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *Workshop of IEEE Conf. Comput. Vis.*, Oct 2019. 1, 3

[20] http://cc0textures.com. Cctextures dataset. In *Creative Commons CC0 1.0 Universal License*. 3

[21] https://3dmodelhaven.com/. Haven dataset. In *Creative Commons CC0 1.0 Universal License*. 3

[22] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[23] Mona Jalal, Josef Spjut, Ben Boudaoud, and Margrit Betke. Sidod: A synthetic image dataset for 3d object pose recognition with distractors. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 475–477, 2019. 8

[24] Donghyun Kim, Kaihong Wang, Stan Sclaroff, and Kate Saenko. A broad study of pre-training for domain generalization and adaptation, 2022. 2

[25] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for

fine-grained recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 301–320, 2016. 2

[26] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *Proc. Int. Conf. on Learn. Rep.*, 2022. 1, 2, 6, 7

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 2

[28] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *Proc. Int. Conf. Mach. Learn.*, volume 97, pages 4013–4022, 2019. 3

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3431–3440, 2015. 1

[30] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22:1345–1359, 2010. 1

[31] Y. Pan, T. Yao, Y. Li, Y. Wang, C. Ngo, and T. Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2234–2242, 2019. 3

[32] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Workshop of IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018. 1, 2, 3, 8

[33] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255, 2019. 3

[34] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020. 2

[35] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2

[36] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2021. 3, 8

[37] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3234–3243, 2016. 1, 3

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115:211–252, 2015. 1

[39] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. In *Proc. Int. Conf. on Learn. Rep.*, 2022. 3

[40] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3723–3732, 2018. 2, 3, 6

[41] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. 2, 4

[42] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6, 2019. 2

[43] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proc. IEEE Int. Conf. Comput. Vis.*, Oct 2017. 2

[44] Zeren Sun, Yazhou Yao, Xiu-Shen Wei, Yongshun Zhang, Fumin Shen, Jianxin Wu, Jian Zhang, and Heng Tao Shen. Webly supervised fine-grained recognition: Benchmark datasets and an approach. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 10602–10611, October 2021. 2

[45] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 8725–8735, 2020. 2, 3, 6, 8

[46] Hui Tang and Kui Jia. Vicinal and categorical domain adaptation. *Pattern Recognition*, 115, 2021. 3

[47] Hui Tang, Yaowei Wang, and Kui Jia. Unsupervised domain adaptation via distilled discriminative clustering. *Pattern Recognition*, 127:108638, 2022. 2, 3, 6, 8

[48] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017. 1, 2, 3

[49] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *Proc. Neur. Info. Proc. Sys.*, volume 34, pages 24261–24272, 2021. 2, 4

[50] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Workshop of IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2018. 3

[51] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journ. of Mach. Learn. Res.*, 9:2579–2605, 2008. 8

[52] VSR Veeravasarapu, Constantin Rothkopf, and Ramesh Visvanathan. Adversarially tuned scene generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[53] VSR Veeravasarapu, Constantin Rothkopf, and Ramesh Visvanathan. Model-driven simulations for computer vision. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1063–1071, 2017. 3

[54] V. S. R. Veeravasarapu, Rudra Narayan Hota, Constantin A. Rothkopf, and Visvanathan Ramesh. Simulations for validation of vision systems. *CoRR*, abs/1512.01030, 2015. 3

[55] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In Zhi-Hua Zhou, editor, *Proc. Int. Jo. Conf. of Artif. Intell.*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track. 2

[56] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), Jul 2020. 2, 3

[57] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5026–5035, 2019. 3