Appendix for
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

The catalog of this appendix is in the following.

- Sec. A summarizes our main contributions.
- Sec. B provides a well-organized summary for the paper novelty.
- Sec. C makes more clarifications on our empirical study.
- Sec. D.1 examines the learning process, Sec. D.2 visualizes the saliency map, and Sec. D.3 depicts more impact of data augmentations for the comprehensive comparison between fixed-dataset periodic training and training on non-repetitive samples.
- Sec. E evaluates various network architectures by plotting their learning curves.
- Sec. F shows the learning process of different pre-training data using domain adaptation as the downstream task.
- Sec. G presents more details on our proposed S2RDA benchmark, e.g., comparing synthetic data with real data from different angles.
- Sec. H provides other implementation details for supervised learning/pre-training and downstream domain adaptation.
- Sec. I reviews other related works on real datasets, data manipulation [55,62,66], deep models [15,27,63], transfer learning [25, 45, 73], domain adaptation [68], and OOD generalization [66].

A. Our Main Contributions

Our main contributions are summarized as follows.

- On the well-controlled IID experimental condition enabled by 3D rendering, we empirically verify the typical insights on shortcut learning, PAC generalization, and variance-bias trade-off, and explore the effects of changing data regimes and network structures on model generalization. The key design wherein is to compare the traditional fixed-dataset periodic training with a new strategy of training on non-repetitive samples.
- We explore how variation factors of an image affect the model generalization, e.g., object scale, material texture, illumination, camera viewpoint, and background, and in return provide new perceptions for data generation.
- Using the popular simulation-to-real classification adaptation as a downstream task, we investigate how synthetic data pre-training performs by comparing with pre-training on real data. We have some surprising and important discoveries including synthetic data pre-training is also prospective and a promising paradigm of pre-training on big synthetic data together with small real data is proposed for realistic supervised pre-training.
- We propose a more large-scale synthetic-to-real benchmark for classification adaptation (termed S2RDA), on which we also provide a baseline performance analysis for representative DA approaches.

B. Summary of Paper Novelty

Now it is becoming more and more important to work on methods that use simulated data but perform well in practical domains whose data or annotation are difficult to acquire, e.g., medical imaging. However, previous research works *have not* studied various factors on a synthesized dataset for image classification and domain adaptation comprehensively and systematically. To fill the gap, we

*Corresponding author.

present *the first work*, ranging from bare supervised learning to downstream domain adaptation. It provides many new, valuable learning insights for OOD/real data generalization, though the verification of some existing, known theories in our well-controlled IID experimental condition has also been done for comprehensive coverage. It is essential for synthetic data learning analysis, which is *completely missing* in the context of image classification. We clarify the paper novelty below.

- The motivation that we utilize synthetic data to verify typical theories and expose new findings is novel. Real data are noisy and uncontrolled, which may hinder the verification of typical theories and exposure to new findings. In the context of image classification, existing works verify classical theories and reveal new findings on real data. However, the process of acquiring real data cannot be controlled, the annotation accuracy cannot be guaranteed, and there may be duplicate images in the training set and test set, which leads to the fact that the training set and test set are no longer independent and identically distributed (IID). To remedy them, we resort to synthetic data generated by 3D rendering with domain randomization.

- The comparison between fixed-dataset periodic training and training on non-repetitive samples and the study of shortcut learning on our synthesized dataset are novel. We admit that some of our findings are classical theories, e.g., PAC generalization and variance-bias trade-off, which should be verified when one introduces a new dataset. We introduce a new dataset of synthetic data and thus do such a study for comprehensive coverage, which first compares fixed-dataset periodic training with training on non-repetitive samples generated by 3D rendering. Particularly, we also verify a recent, significant perspective of shortcut learning and design new experiments to demonstrate that randomizing the variation factors of training images can block shortcut solutions that rely on context clues in the background.

- Investigating the learning characteristics and properties of our synthesized new dataset comprehensively is novel, and our experiments yield many interesting and valuable observations. Synthetic data are cheap, label-rich, and well-controlled, but there hasn't been a comprehensive study of bare supervised learning on synthetic data in the context of image classification. To our knowledge, we are the first to investigate the learning characteristics and properties of our synthesized new dataset comprehensively, in terms of refreshed architecture, model capacity, training data quantity, data augmentation, and rendering variations. The empirical study on bare supervised learning yields many new findings, e.g.,

  - IID and OOD generalizations are some type of zero-sum game,
  - ViT performs surprisingly poorly,
  - there is always a bottleneck from synthetic data to OOD/real data,
  - neural architecture search (NAS) should also consider the search for data augmentation, and
  - different factors and even their different values have uneven importance to IID generalization.

- Synthetic data pre-training, its comparison to real data pre-training, and its application to downstream synthetic-to-real classification adaptation are novel, and our experiments yield many interesting and valuable observations. To our knowledge, there is little research on pre-training for domain adaptation. Kin et al. [29] preliminarily study the effects of real data pre-training on domain transfer tasks. Differently, we focus on the learning utility of synthetic data and take the first step towards clearing the cloud of mystery surrounding how different pre-training schemes including synthetic data pre-training affect the practical, large-scale synthetic-to-real classification adaptation. Besides, we first study and compare pre-training on the latest MetaShift dataset [40]. The empirical study on downstream domain adaptation yields many new findings, e.g.,

  - DA fails without pre-training,
  - 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 is better than pre-training on real data (e.g., ImageNet) in our study, and
  - Big Synthesis Small Real is worth deeply researching.

- Our introduced S2RDA benchmark is novel and will advance the field of domain adaptation research on transfer from synthetic to real.

Our findings may challenge some of the current conclusions, but they also shed some light on the important fields in computer vision and take a step towards uncovering the mystery of deep learning.

Table A1. Domain adaptation performance on SubVisDA-10 with varied pre-training schemes (ResNet-50). Green or red: Best Acc. or Mean in each row (among compared DA methods).

Pre-training Data	# Iters	# Epochs	No Adaptation		DANN		MCD		RCA		SRDC		DisClusterDA	
			Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean	Acc.	Mean
ImageNet-990	200K	10	50.11	45.45	55.68	54.67	58.84	58.44	58.79	60.18	60.25	57.68	57.62	57.42
ImageNet-990+Ours	200K	9	52.87	**48.85**	**58.42**	**58.02**	60.52	**62.27**	**62.28**	**63.35**	**63.60**	**60.89**	**61.90**	**63.10**
ImageNet	200K	10	**53.24**	45.38	57.77	55.59	**61.90**	61.75	61.59	60.72	62.56	59.24	61.18	59.01

C. More Clarifications on Our Empirical Study

Necessity of domain randomization study in Table 2. We expect that assessing variation factors should be essential for synthesizing data for image classification, which is *missing* in previous work [48]. We admit that such a study has been done for detection and scene understanding in prior methods [49, 62, 64], but the generalizability of those results to image classification is *lack of guarantee*. Thus, we follow them and do the study for image classification, where we consider a *different* set of factors (e.g., light, texture, and flying distractors in [64]). It may be expected that a fixed value would underperform randomized values, but how much each factor and each value degrade is *unknown*. One cannot know how important they are without such a *quantitative* study. Insightfully, our new results in Table 2 also suggest that the under-explored direction of weighted rendering is worth studying and provide preliminary guidance/prior knowledge for learning factor distribution.

Additonal experiments on ImageNet-990+Ours improving over ImageNet-990. To further justify that ImageNet-990+Ours improves over ImageNet-990, we do additional experiments by using a different cosine decay schedule for the learning rate: $\eta_p = \eta_1 + 0.5(\eta_0 - \eta_1)(1 + \cos(\pi p))$, where p is the process of training iterations normalized to be in $[0, 1]$, the initial learning rate $\eta_0 = 0.1$, and the final learning rate $\eta_1 = 0.001$. The results for several DA methods are reported in Table A1. As we can see, with fine-grained subclasses merged into one, ImageNet-990 underperforms ImageNet by a large margin, suggesting that it may be helpful to use fine-grained visual categorization for pre-training; in contrast, by adding our 120K synthetic images, ImageNet-990+Ours is comparable to or better than ImageNet, confirming the utility of our synthetic data.

Pre-training with an increased number of classes helps DA. In Table 3, we have compared SubImageNet involving only target classes in training with ImageNet involving both target and non-target classes; with abundant pre-training epochs, the latter is evidently better than the former, indicating that learning rich category relationships is helpful for downstream DA. *A similar phenomenon is observed in synthetic data pre-training.* For example, we have done experiments of pre-training on the synthetic domain of our proposed S2RDA-49 task; compared with pre-training on Ours (120K images, 10 classes), MCD improves by 5.97% in Acc. and 6.06% in Mean, and DisClusterDA improves by 4.69% in Acc. and 5.36% in Mean.

Necessity and applicability of the proposed S2RDA. Note that SRDC outperforms the baseline No Adaptation by $\sim$10% on S2RDA-49 and DisClusterDA outperforms that by $\sim$5%, verifying the efficacy of these DA methods. The observations also demonstrate that S2RDA *can* benchmark different DA methods. Compared to SubVisDA-10 (cf. Table 3), SRDC degrades by $\sim$7% on S2RDA-49, which is reasonable as our real domain contains more practical images from real-world sources, though *our synthetic data contain much more diversity, e.g., background* (cf. Fig. 1). Differently, S2RDA-MS-39, which decreases by $>$20% over S2RDA-49, evaluates different DA approaches on the *worst/extreme cases* (cf. Fig. 6), making a more comprehensive comparison and acting as a touchstone to examine and advance DA algorithms. Reducing the domain gap between simple and difficult backgrounds is by nature one of the key issues in simulation-to-real transfer, as also shown in [48]; therefore, reducing such a domain gap is one of the criteria for judging excellent DA methods. To sum up, S2RDA is a *better* benchmark than VisDA-2017, as it has more realistic synthetic data and more practical real data with more object categories, and enables a larger room of improvement for promoting the progress of DA algorithms and models.

D. Fixed-Dataset Periodic Training vs. Training on Non-Repetitive Samples

D.1. Examining Learning Process

In Fig. A1, we examine the learning process of fixed-dataset periodic training and training on non-repetitive samples based on ResNet-50 with no, weak, and strong data augmentations. To this end, we plot the evolving curves of the following eight quantities with the training: training loss measured on the synthetic training set, test loss (IID) measured on the synthetic IID test set, training accuracy measured on the synthetic training set, test accuracy (IID) mea-

sured on the synthetic IID test set, test loss (IID w/o BG) measured on the synthetic IID without background test set, test loss (OOD) measured on the SubVisDA-10 real/OOD test set, test accuracy (IID w/o BG) measured on the synthetic IID without background test set, and test accuracy (OOD) measured on the SubVisDA-10 real/OOD test set. The accuracy is measured using the ground truth labels, just for visualization.

D.2. Visualizing Saliency Map

We visualize the saliency maps, obtained from the ResNet-50 (Fig. A2), ViT-B (Fig. A3), and Mixer-B (Fig. A5) trained on a fixed dataset or non-repetitive samples with no data augmentation. We consider two types of saliency visualization methods: input gradients which backpropagates the output score at the ground-truth category to the input image, and Grad-CAM which weights the feature maps with the gradients w.r.t. the features. For ViT-B, in Fig. A4, we also visualize the attention maps of the classification token to all image patches at the last multi-head self-attention layer. The last five columns correspond to results at the 20-th, 200-th, 2K-th, 20K-th, and 200K-th training iterations respectively. The example image in each row is randomly selected from IID test data.

D.3. More Impact of Data Augmentations

From Table 1, in OOD tests, training on non-repetitive images with no augmentation is superior to the fixed-dataset periodic training with weak augmentation, but far inferior to that with strong augmentation. It to some extent implies that the image transformations produced by 3D rendering itself do contain the hand-crafted weak augmentation changing pixel position in an image, but not the strong one changing both position and value of pixels.

E. Evaluating Various Network Architectures

In Fig. A6, we evaluate various network architectures by plotting their learning curves, in terms of training loss, test loss (IID), training accuracy, test accuracy (IID), test loss (IID w/o BG), test loss (OOD), test accuracy (IID w/o BG), and test accuracy (OOD). Various network architectures — ResNet-50 [23], ViT-B [15], and Mixer-B [63] are trained on non-repetitive samples with strong data augmentation.

We note that ViT performs poorly despite data augmentation and more training epochs. It may be because ViT cannot well fit datasets with high dimension and large variance, and does not learn features less dependent on the background. Two possible solutions to improve are decreasing the input patch size for fine-grained feature interaction and smoothing the dataset (e.g. mixup [71]) for data distribution completeness.

F. Comparing Pre-training for Domain Adaptation

In Fig. A7, we compare different pre-training data using domain adaptation (DA) on SubVisDA-10 as the downstream task and show the learning process for several representative DA approaches. The considered pre-training schemes include **(1)** No Pre-training where the model parameters are randomly initialized, **(2)** Ours denotes our synthesized 120K images of the 10 object classes shared by SubVisDA-10, **(3)** SubImageNet is the subset collecting examples of the 10 classes from ImageNet [13], **(4)** ImageNet (10 Epoch) has 1K classes and 10 pre-training epochs, and **(5)** ImageNet$^\bigstar$ uses the official ResNet-50 [23] checkpoint pre-trained on ImageNet for 120 epochs. The compared DA methods include No Adaptation that trains the model only on the labeled source data, DANN [18], MCD, [53], RCA [10], SRDC [59], and DisClusterDA [61].

G. More Details on Our Proposed S2RDA Benchmark

Dataset Details. Our proposed Synthetic-to-Real (S2RDA) benchmark for more practical visual domain adaptation (DA) includes two challenging transfer tasks of S2RDA-49 and S2RDA-MS-39. In Fig. A8, we show the distribution of the number of images per class in each real domain, which is exhibited to be a long-tailed distribution where a small number of classes dominate. How we collect the real data from diverse real-world sources is recorded in respective files included in the code. Our S2RDA dataset is publicly available at https://pan.baidu.com/s/1fHHaqrEHbUZLXEg9XKpgSg?pwd=w9wa.

Comparing Synthetic Data with Real Data. We provide quantitative and qualitative comparisons for VisDA-2017, our synthesized dataset, and ImageNet as follows. The mean and standard deviation of VisDA-2017, our synthesized dataset, and ImageNet are [0.878, 0.876, 0.874] and [0.207, 0.210, 0.216], [0.487, 0.450, 0.462] and [0.237, 0.251, 0.270], and [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225] respectively. As we can see, the statistics of our synthesized dataset are closer to the real dataset ImageNet than VisDA-2017. It is consistent with the observation in Table 1 that our synthesized dataset used for training yields higher OOD/real test accuracy than SubVisDA-10. Except for quantitative comparisons, we have also provided the qualitative visualization for the three datasets and the proposed S2RDA benchmark in Fig. 1 and Fig. 6 respectively, which demonstrates that our synthesized dataset is visually more similar to ImageNet.

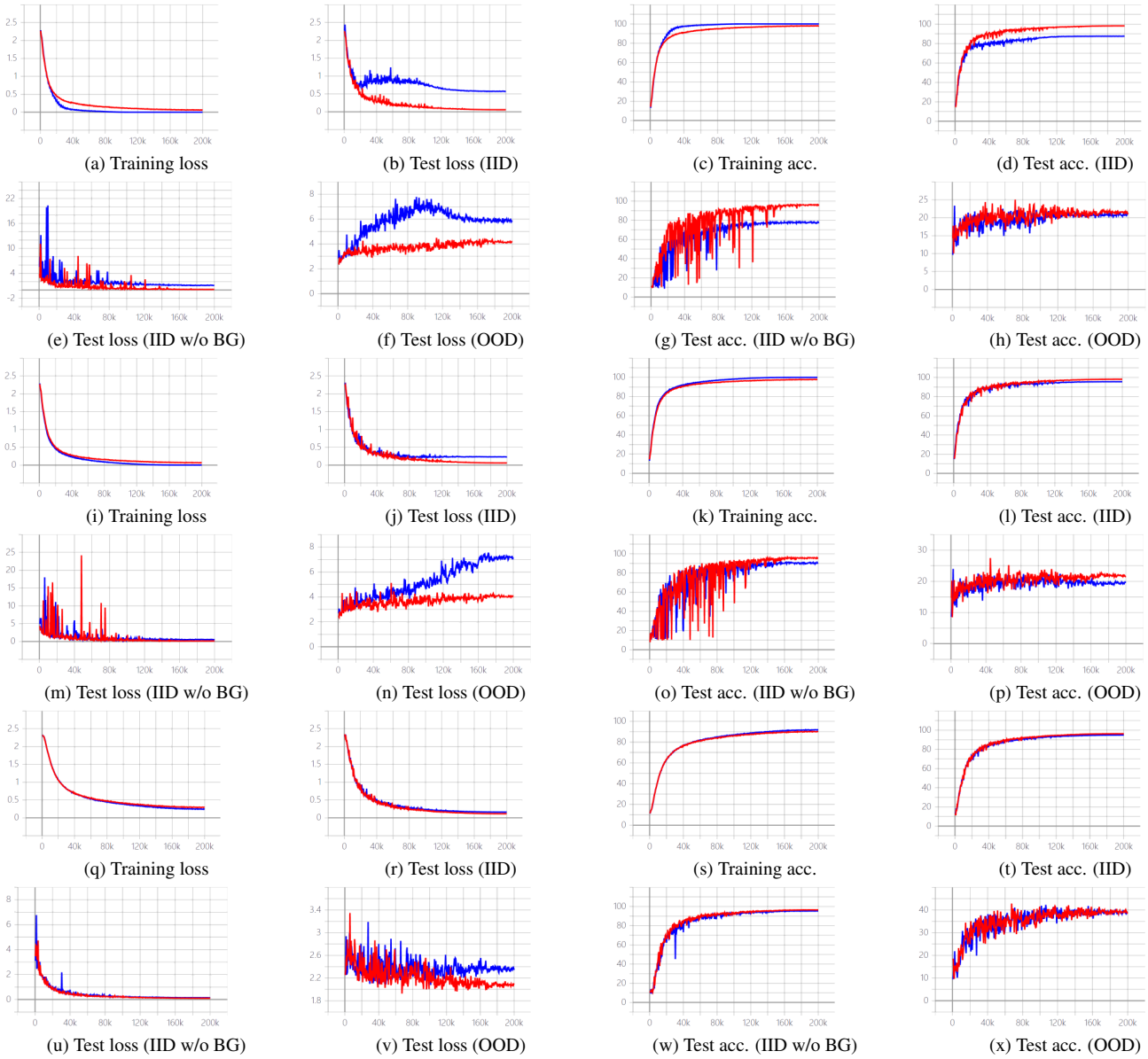
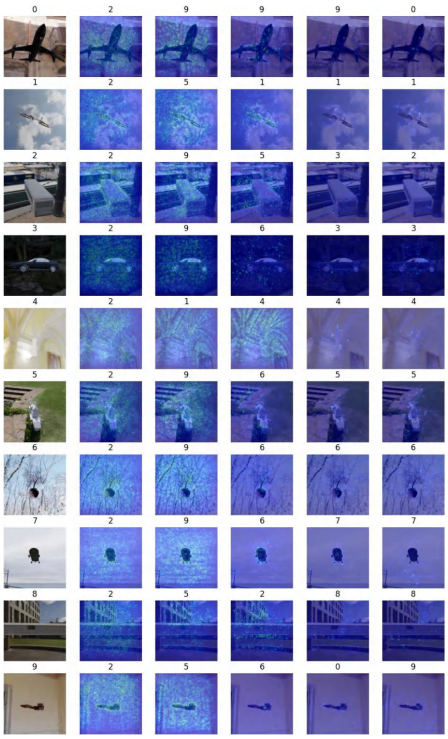


Figure A1. Learning process of training ResNet-50 on a fixed dataset (blue) or non-repetitive samples (red). Note that (a-h), (i-p), and (q-x) are for no, weak, and strong data augmentations respectively.

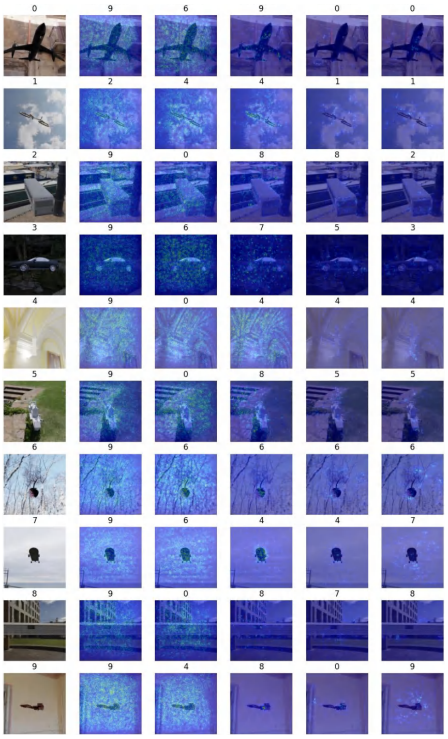
H. Other Implementation Details

Supervised Learning/Pre-training. For each backbone in Sec. 4.1, all its layers up to the second last one are used as the feature extractor and the neuron number of its last FC layer is set as 10 to have the classifier. We use the cosine learning rate schedule: the learning rate is adjusted by $\eta_p = 0.5\eta_0(1+\cos(\pi p))$, where p is the process of training iterations normalized to be in $[0, 1]$ and the initial learning rate $\eta_0 = 0.01$. The momentum, weight decay, and random seed are set as 0.9, 0.0001, and 1 respectively. In light of fairness, the final normalization operation uses the ImageNet statistics consistently for all experiments.

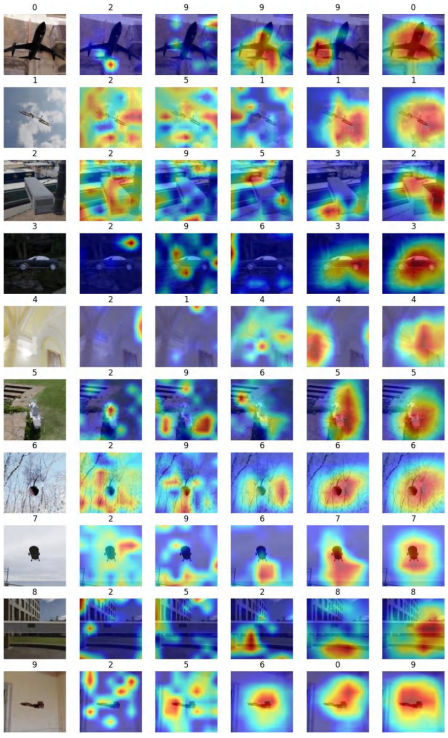
Downstream Domain Adaptation. In domain adaptation training, we use all labeled source samples and all unlabeled target samples as the training data. In each base model, the last FC layer is replaced with a new task-specific FC layer as the classifier. We fine-tune the pre-trained layers and train the new layer from scratch, where the learning rate of the latter is 10 times that of the former. The learning rate is adjusted by $\eta_p = \eta_0(1 + \alpha p)^{-\beta}$, where p denotes the training process of training epochs normalized to



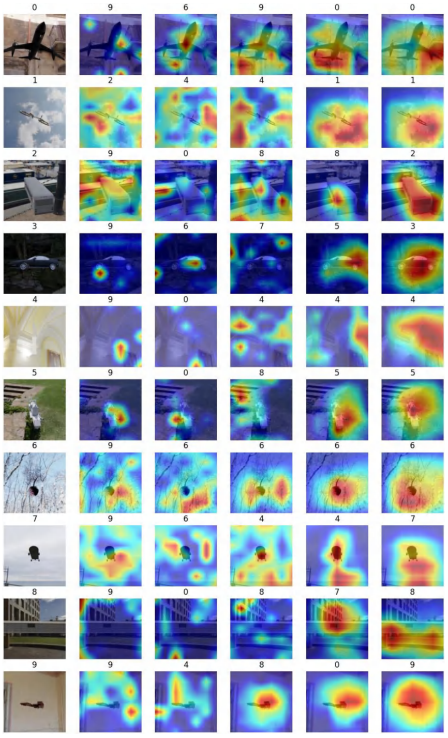
(a) Fixed



(b) Non-repetitive

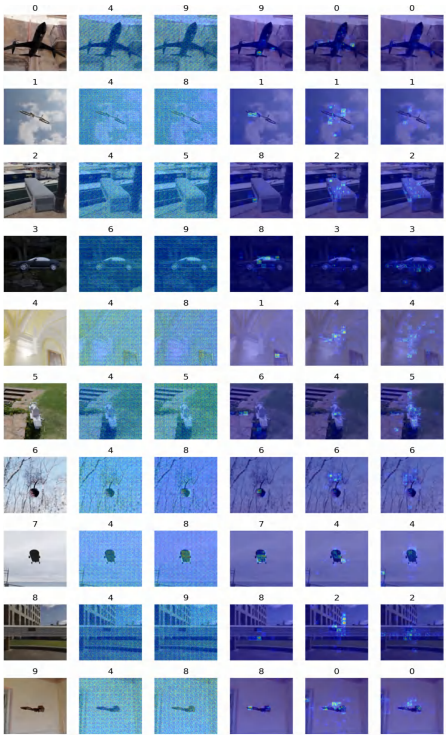


(c) Fixed

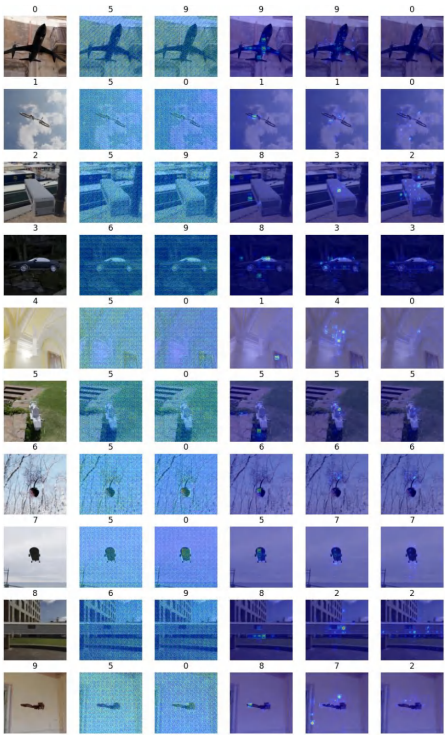


(d) Non-repetitive

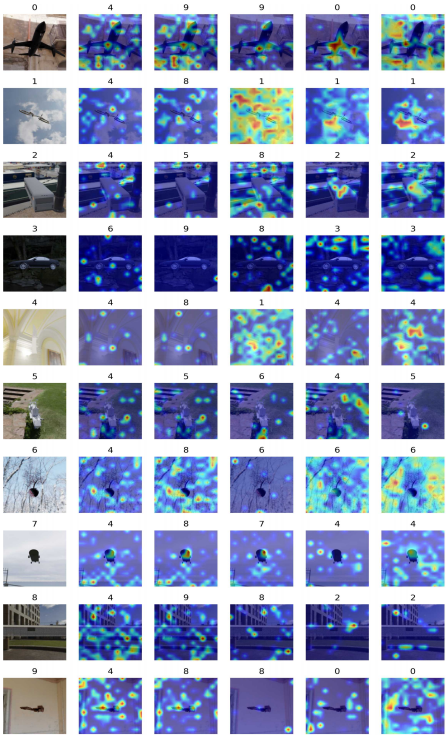
Figure A2. Saliency maps of randomly selected IID test samples, obtained from the ResNet-50 trained on a fixed dataset or non-repetitive samples with no data augmentation, at the 20-th, 200-th, 2K-th, 20K-th, and 200K-th training iterations. Note that rows 1 and 2 show input gradients [56] and gradient weighted class activation maps [54] on input images respectively; the number on top of each picture means the ground-truth (first column) or predicted labels (other columns).



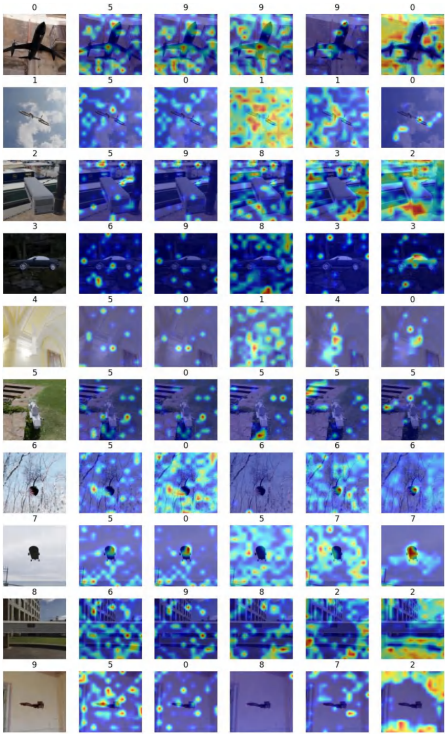
(a) Fixed



(b) Non-repetitive

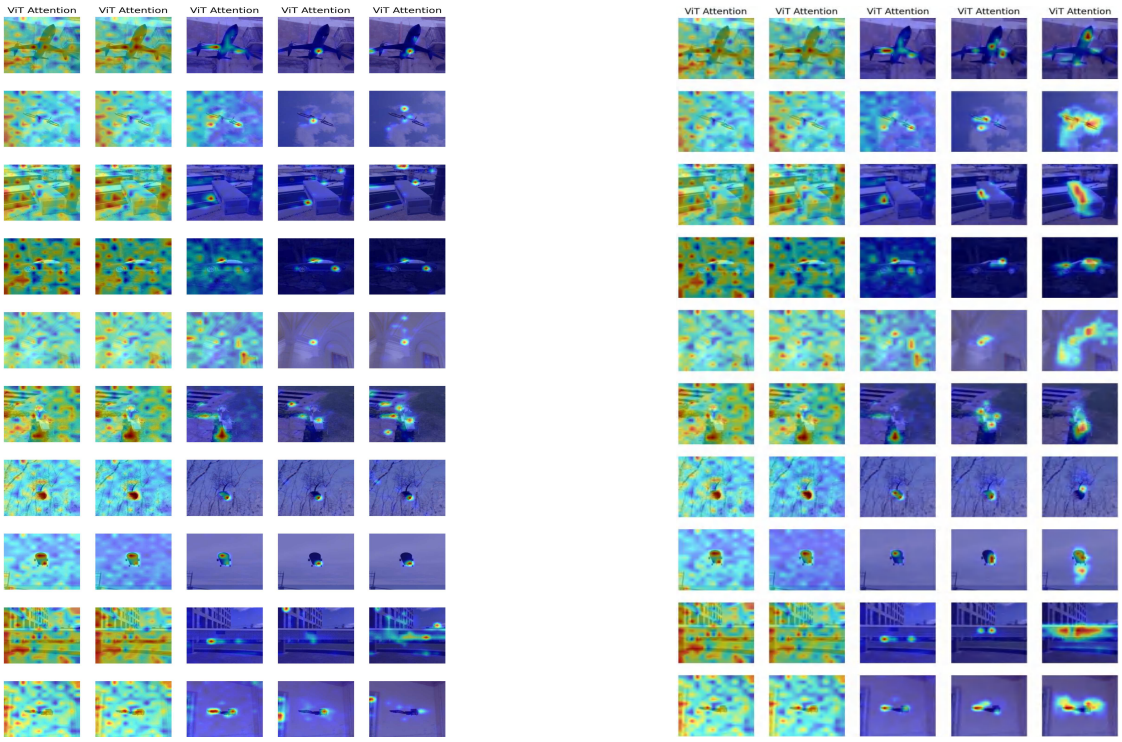


(c) Fixed



(d) Non-repetitive

Figure A3. Saliency maps of randomly selected IID test samples, obtained from the ViT-B trained on a fixed dataset or non-repetitive samples with no data augmentation, at the 20-th, 200-th, 2K-th, 20K-th, and 200K-th training iterations. Note that rows 1 and 2 show input gradients [56] and gradient weighted class activation maps [54] on input images respectively; the number on top of each picture means the ground-truth (first column) or predicted labels (other columns).



(a) Fixed (b) Non-repetitive

Figure A4. Attention maps [1] of randomly selected IID test samples, obtained from the ViT-B trained on a fixed dataset or non-repetitive samples with no data augmentation, at the 20-th, 200-th, 2K-th, 20K-th, and 200K-th training iterations.

be in $[0, 1]$, the initial learning rate η_0 is 0.001 for MCD and 0.0001 for other methods, $\alpha = 10$, and $\beta = 0.75$. The momentum, weight decay, and random seed are set as 0.9, 0.0001, and 0 respectively. By convention, strong and weak data augmentations are applied in pre-training and domain adaptation respectively. For domain adaptation on our proposed S2RDA benchmark, we use ResNet-50 as the backbone, which is initialized by the official ImageNet pre-trained checkpoint [23]. The initial learning rate is set as 0.0001 across all experiments. Other implementation details are the same as those described above.

More details are as follows.

1. For 3D rendering with domain randomization, the monocular camera default in BlenderProc [14] is used in the 3D renderer.

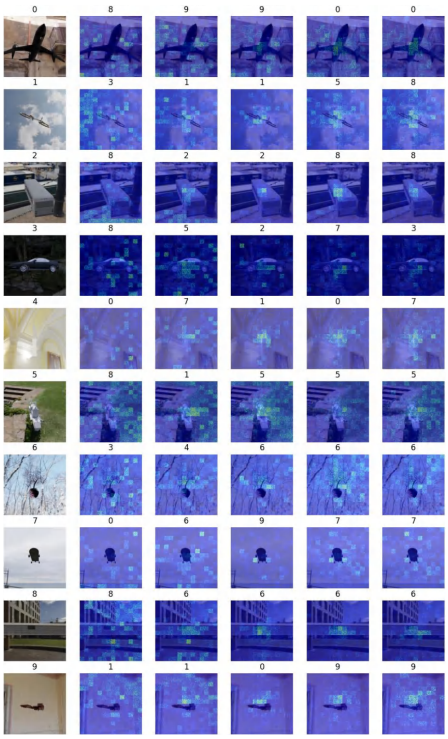
2. SubVisDA-10 includes the following 10 classes: airplane, bicycle, bus, car, knife, motorbike, plant, skateboard, train, and truck.

3. MetaShift [40] we use is a filtered version of 2559865 images from 376 classes by running the officially provided code of dataset construction. It is formed by setting a threshold for subset size ($>= 25$) and subset number in one class (> 5).

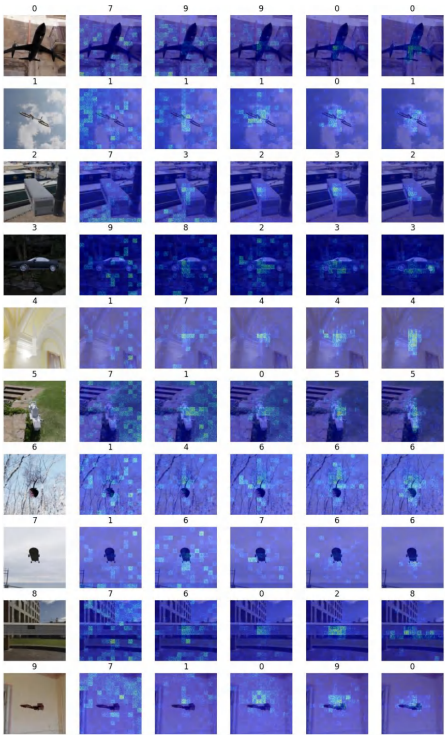
4. Mean class precision is the average over recognition precisions of all classes. It is an indicator of class imbalance that different categories have different prediction accuracy. When it deviates from the overall accuracy in a test, class imbalance happens.

5. In Table 3, the number highlighted by the green color indicates the best Acc. in each row (among all compared DA methods), the number underlined by the red color indicates the best Mean in each row (among all compared DA methods), and the bold number in each column indicates the best result among all considered pre-training schemes. In Table 4, the bold number highlighted by the green color indicates the best Acc. in each row (among all compared DA methods), and the bold number underlined by the red color indicates the best Mean in each row (among all compared DA methods).

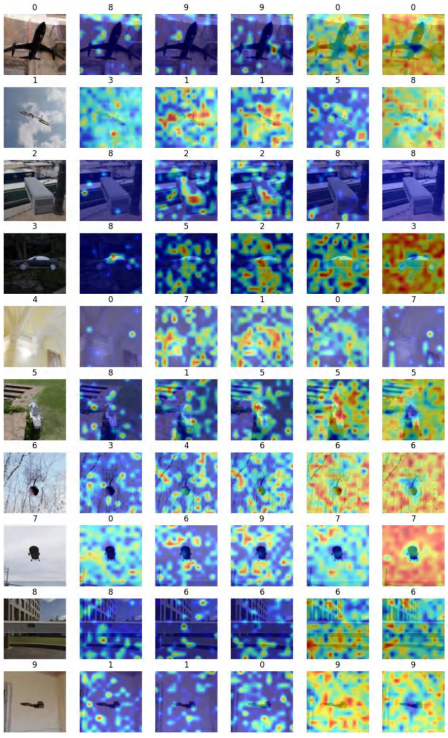
6. PyTorch [47] is used for implementation. Grid Search is used for hyperparameter tuning. We use an 8-GPU NVIDIA GeForce GTX 1080 and an 8-GPU NVIDIA Tesla M40 to run experiments. For the used assets, we have cited the corresponding references in the main paper, and we mention their licenses here: CCTextures under CC0 License, Haven under CC0 License, ShapeNet under a custom license, VisDA-2017 under



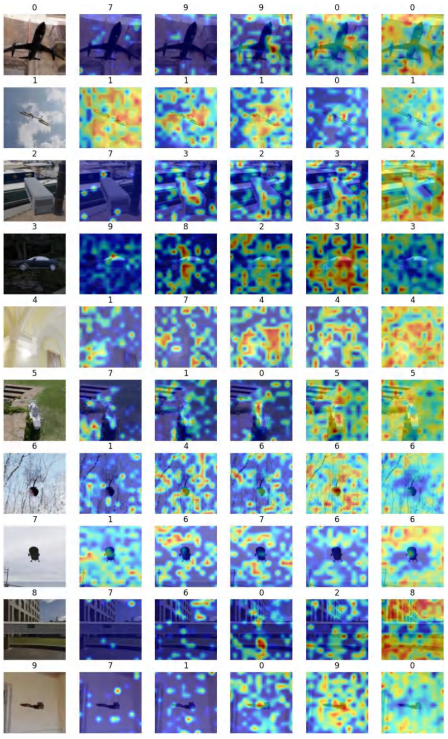
(a) Fixed



(b) Non-repetitive

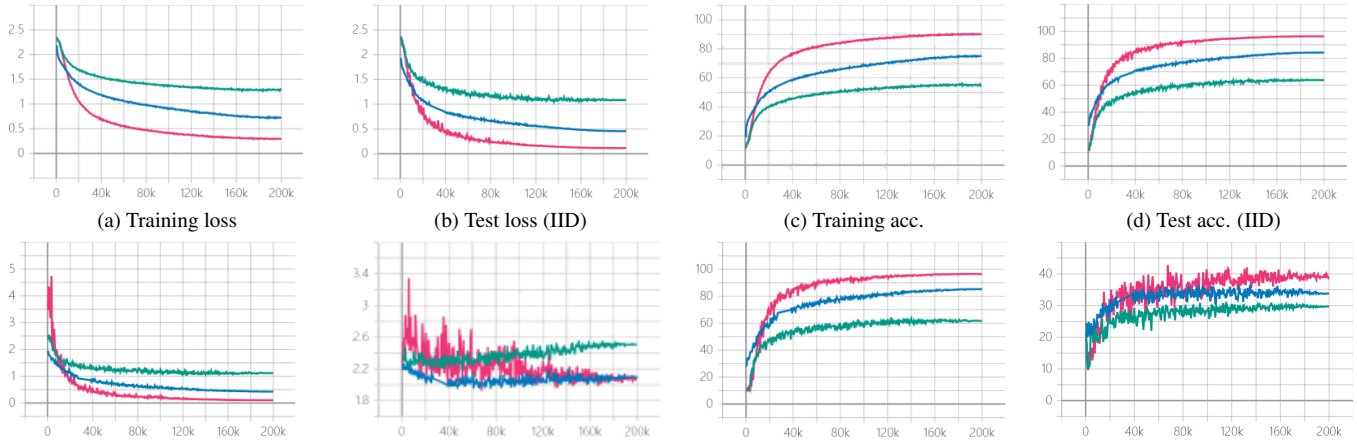


(c) Fixed



(d) Non-repetitive

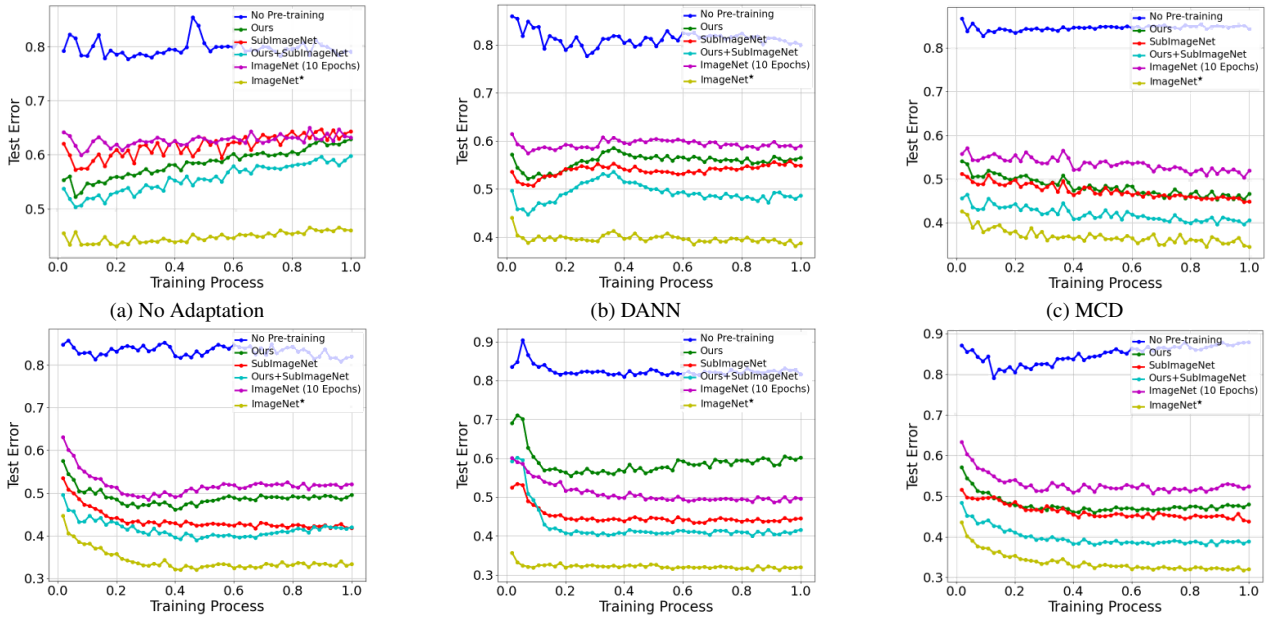
Figure A5. Saliency maps of randomly selected IID test samples, obtained from the Mixer-B trained on a fixed dataset or non-repetitive samples with no data augmentation, at the 20-th, 200-th, 2K-th, 20K-th, and 200K-th training iterations. Note that rows 1 and 2 show input gradients [56] and gradient weighted class activation maps [54] on input images respectively; the number on top of each picture means the ground-truth (first column) or predicted labels (other columns).



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

(e) Test loss (IID w/o BG) (f) Test loss (OOD) (g) Test acc. (IID w/o BG) (h) Test acc. (OOD)

Figure A6. Learning process of training various network architectures on non-repetitive samples with strong data augmentation. Note that red, green, and blue indicate ResNet-50, ViT-B, and Mixer-B respectively.



(a) No Adaptation (b) DANN (c) MCD

(d) RCA (e) SRDC (f) DisClusterDA

Figure A7. Learning process (Acc.) of domain adaptation when varying the pre-training scheme.

MIT License, ImageNet under a custom license, and MetaShift under MIT License.

I. Other Related Works

Real Datasets. A lot of large-scale real datasets [13, 19, 32, 40, 41, 51, 57, 58] have harnessed and organized the explosive image data from Internet or the real world for deep learning of meaningful visual representations. For example, ImageNet [13] is a large-scale database of images built upon the backbone of the WordNet structure; ImageNet-1K, consisting of 1.28M images from 1K common object categories, which serves as the primary dataset for pre-training deep models for computer vision tasks. Barbu et al. [4] collect a large real-world test set for more realistic object recognition, ObjectNet, which has 50K images and is bias-controlled. Ridnik et al. [51] dedicatedly preprocess the full set of ImageNet — ImageNet-21K with the WordNet hierarchical structure utilized, such that high-quality efficient pre-training on the resulted ImageNet-21K-P (of 12M images) can be made for practical use. In [57], JFT-300M of more than 375M noisy labels for 300M images is exploited to study the effects of pre-training on current vision tasks.

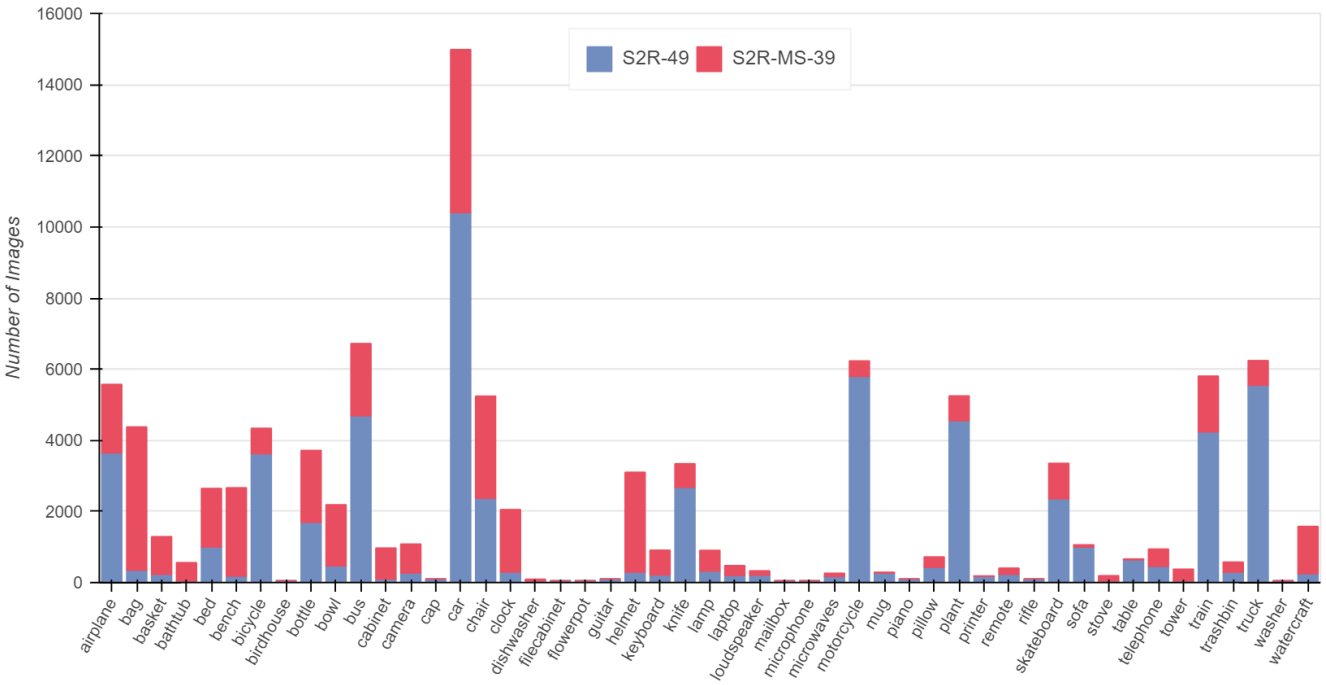


Figure A8. Distribution of the number of images per class in each real domain of our proposed S2RDA.

Geirhos et al. [19] construct StylizedImageNet by replacing the object texture in an image with a random painting style via style transfer, to learn a shape-based representation. MS COCO [41] has 328K images, where considerably more object instances exist as compared to ImageNet, enabling deep models to learn precise 2D localization. MetaShift [40] of 2.56M natural images ($\sim$ 400 classes) is formed by context guided clustering of the images from GQA [24], a cleaned version of Visual Genome [33] connecting language and vision. CCT-20 [5] and PACS [36] are designed to measure recognition generalization to novel visual domains. Some works [32, 58] leverage free, numerous web data to benchmark or assist fine-grained recognition. Some small datasets are used as benchmarks for semi-supervised learning [11, 34, 44] or domain generalization [5, 36].

Data Manipulation. Deep models are hungry for more training data in that the generalization ability often relies on the quantity and diversity of training samples. To improve model generalization with a limited set of training data available, the cheapest and simplest way is data manipulation, which increases the sample diversity from two different perspectives of data augmentation and data generation. The former applies a series of random image transformations [55] or appends adversarial examples at each iteration [65]; the latter uses generative models to generate diverse and rich data such as Variational Auto-Encoder (VAE) [50] and Generative Adversarial Network (GAN) [20] or renders 3D object models into RGB images via domain randomization [49, 62, 64].

Deep Models. Deep model has strong representational capacity in that they can learn powerful, hierarchical representations when trained on large amounts of data; they can be highly scalable from various aspects of architectural innovation, such as spatial exploitation, depth, multi-path, width, feature-map exploitation, channel boosting, and attention. Deep Convolutional Neural Networks (CNNs) have been popularized for decades in a wide range of computer vision tasks. A comprehensive survey for CNNs can be found in [27]. The most commonly used CNN-based network architecture is ResNet [23], which reformulates the layers as learning residual functions concerning the layer inputs, instead of learning unreferenced functions. Recently, some new types of network architectures have emerged, such as ViT [15] and MLP-Mixer [63]. ViT stacks a certain number of multi-head self-attention layers and is applied directly to sequences of fixed-size image patches. MLP-Mixer is based exclusively on multi-layer perceptrons (MLPs) and contains two types of MLP layers: one for mixing channels in individual image patches and one for mixing features across patches of different spatial locations. In this work, we experiment on the three representative types of networks.

Transfer Learning. There has been a huge literature in the field of transfer learning [25,45,73], where the paradigm of pre-training and then fine-tuning has made outstanding achievements in many deep learning applications. Extensive studies have been done for supervised pre-training [17, 29–31, 67, 69]. For example, Yosinski et al. [69] examine the transferability of features at different layers along the network; the relationship between ImageNet accuracy and transferability is evaluated in [31]; BiT [30] provides a recipe of the minimal number of existing tricks for pre-training and downstream transferring; in [67], an MLP projector is added before the classifier to improve the transferability; LOOK [17] solves the problem of overfitting upstream tasks by only allowing nearest neighbors to share the class label, in order to preserve the intra-class semantic difference; particularly, Kin et al. [29] preliminarily study the effects of pre-training on domain transfer tasks, from the aspects of network architectures, size, pre-training loss, and datasets. Another popular branch of self-supervised learning is increasingly important for transfer learning. Previous works have proposed various pretext tasks, such as image inpainting and jigsaw puzzle [26]. Recent works concentrate on self-supervised/unsupervised pre-training [6–8, 21, 22, 26, 28, 39] and have shown powerful transferability on multiple downstream tasks, comparable to supervised pre-training. They often rely on contrastive learning to learn visual representations of rich intra-class diversity [70], e.g., contrasting feature embeddings [6, 8, 21, 22, 28] or cluster assignments [7] of anchor, positive, and negative instances. Note that CDS [28] proposes a second self-supervised pre-training stage using the unlabeled downstream data from multiple domains, which applies instance discrimination not only in individual domains but also across domains. Also, many researchers are devoted to improving fine-tuning by leveraging the pre-trained ImageNet knowledge [9], using pre-training data for fine-tuning [43], improving regularization and robustness [37], adapting unfamiliar inputs [2], applying the easy two-step strategy of linear probing and then full fine-tuning [35], to name a few. Different from [29], we focus on the utility of synthetic data and take the first step towards clearing the cloud of mystery surrounding how different pre-training schemes including synthetic data pre-training affect the practical, large-scale synthetic-to-real adaptation.

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 [10,18,46,53,60,72], such that the domain-invariant features can be learned and thus the task classifier trained on the labeled source data can well generalize to the unlabeled target domain. DANN [18] aligns the source and target domains as a whole by domain-adversarial training, i.e., reversing the signal from a domain discriminator, but does not utilize the discriminative information from the target domain. MCD [53] minimizes the maximum prediction discrepancy between two task classifiers to learn domain-invariant and class-discriminative features. RCA [10] implements the domain-adversarial training based on a joint domain-category classifier to learn class-level aligned features, i.e., invariant at corresponding classes of the two domains. Differently, works of another emerging strategy [12, 42, 59, 61] take steps towards implicit domain adaptation, without explicit feature alignment that could hurt the intrinsic discriminative structures of target data. SRDC [59] uncovers the intrinsic target discrimination via deep discriminative target clustering in both the output and feature spaces with structural source regularization hinging on the assumption of structural similarity across domains. DisClusterDA [61] proposes a new clustering objective for discriminative clustering of target data with distilled informative source knowledge, based on a robust variant of entropy minimization, a soft Fisher-like criterion, and the cluster ordering via centroid classification. In this work, we consider these representative DA methods for the empirical study, and broader introductions to the rich literature are provided in [52, 68].

OOD Generalization. Out-of-distribution (OOD) generalization, i.e., domain generalization, assumes the access to single or multiple different but related domains and aims to generalize the learned model to an unseen test domain. A detailed review for recent advances in domain generalization is presented in [66], which categorizes the popular algorithms into three classes: data manipulation [50, 65], representation learning [16, 38], and learning strategy [3, 50]. For example, adversarial examples are generated to learn robust models in [65]; a progressive domain expansion subnetwork and a domain-invariant representation learning subnetwork are jointly learned to mutually benefit from each other in [38]; Balaji et al. [3] adopt the meta-learning strategy to learn a regularizer that the model trained on one domain can well generalize to another domain.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. 8

[2] Shengnan An, Yifei Li, Zeqi Lin, Qian Liu, Bei Chen, Qiang Fu, Weizhu Chen, Nanning Zheng, and Jian-Guang Lou. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models, 2022. 12

[3] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Proc. Neur. Info. Proc. Sys.*, volume 31. Curran Associates, Inc., 2018. 12

[4] 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. 10

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 11

[6] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 194–203, June 2021. 12

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. Neur. Info. Proc. Sys.*, Red Hook, NY, USA, 2020. Curran Associates Inc. 12

[8] 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. 12

[9] Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M. Alvarez, Zhangyang Wang, and Anima Anandkumar. Contrastive syn-to-real generalization. In *Proc. Int. Conf. on Learn. Rep.*, 2021. 12

[10] S. Cicek and S. Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1416–1425, 2019. 4, 12

[11] 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, 2011. 11

[12] 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. 12

[13] 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. 4, 10

[14] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019. 8

[15] 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. 1, 4, 11

[16] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8208–8217, June 2021. 12

[17] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. In *Proc. Int. Conf. on Learn. Rep.*, 2022. 12

[18] 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. 4, 12

[19] 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. 10, 11

[20] Kamran Ghasedi, Xiaoqian Wang, Cheng Deng, and Heng Huang. Balanced self-paced learning for generative adversarial clustering network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4386–4395, 2019. 11

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proc. Neur. Info. Proc. Sys.*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 12

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 12

[23] 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. 4, 8, 11

[24] 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. 11

[25] Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. Transferability in deep learning: A survey. *ArXiv*, abs/2201.05867, 2022. 1, 12

[26] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4037–4058, 2021. 12

[27] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.*, 53:5455–5516, 2020. 1, 11

[28] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 9123–9132, October 2021. 12

[29] Donghyun Kim, Kaihong Wang, Stan Sclaroff, and Kate Saenko. A broad study of pre-training for domain generalization and adaptation, 2022. 2, 12

[30] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Proc. Eur. Conf. Comput. Vis.*, pages 491–507, 2020. 12

[31] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2019. 12

[32] 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. 10, 11

[33] R. Krishna, Y. Zhu, and O. et al Groth. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, page 32–73, 2017. 11

[34] A. Krizhevsky. Learning multiple layers of features from tiny images. In *Technical report*, 2009. 11

[35] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *Proc. Int. Conf. on Learn. Rep.*, 2022. 12

[36] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, broader and artier domain generalization. In *International Conference on Computer Vision*, 2017. 11

[37] Dongyue Li and Hongyang Zhang. Improved regularization and robustness for fine-tuning in neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Proc. Neur. Info. Proc. Sys.*, volume 34, pages 27249–27262. Curran Associates, Inc., 2021. 12

[38] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 224–233, June 2021. 12

[39] Suichan Li, Dongdong Chen, Yinpeng Chen, Lu Yuan, Lei Zhang, Qi Chu, Bin Liu, and Nenghai Yu. Improve unsupervised pretraining for few-label transfer. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 10201–10210, October 2021. 12

[40] 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. 2, 8, 10, 11

[41] 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. 10, 11

[42] 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. 12

[43] Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Antoni B. Chan, and Rong Jin. Improved fine-tuning by leveraging pre-training data: Theory and practice. *CoRR*, abs/2111.12292, 2021. 12

[44] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *Workshop of Proc. Neur. Info. Proc. Sys.*, 2011. 11

[45] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22:1345–1359, 2010. 1, 12

[46] 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. 12

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Workshop of Proc. Neur. Info. Proc. Sys.*, 2017. 8

[48] 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. 3

[49] 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, 11

[50] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020. 11, 12

[51] 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. 10

[52] 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 un-

supervised adaptation. In *Proc. Int. Conf. on Learn. Rep.*, 2022. 12

[53] 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. 4, 12

[54] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 618–626, 2017. 6, 7, 9

[55] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6, 2019. 1, 11

[56] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014. 6, 7, 9

[57] 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. 10

[58] 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. 10, 11

[59] 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. 4, 12

[60] Hui Tang and Kui Jia. Vicinal and categorical domain adaptation. *Pattern Recognition*, 115, 2021. 12

[61] Hui Tang, Yaowei Wang, and Kui Jia. Unsupervised domain adaptation via distilled discriminative clustering. *Pattern Recognition*, 127:108638, 2022. 4, 12

[62] 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, 3, 11

[63] 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. 1, 4, 11

[64] 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, 11

[65] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Proc. Neur. Info. Proc. Sys.*, volume 31, 2018. 11, 12

[66] 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. 1, 12

[67] Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an MLP perspective. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022. 12

[68] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), Jul 2020. 1, 12

[69] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proc. Neur. Info. Proc. Sys.*, pages 3320–3328, 2014. 12

[70] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Proc. Neur. Info. Proc. Sys.*, 2020. 12

[71] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. Int. Conf. on Learn. Rep.*, 2018. 4

[72] 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. 12

[73] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. 1, 12