JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

South China University of Technology

DEXFORCE
跨维智能

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

Authors: Hui Tang and Kui Jia

eehuitang@mail.scut.edu.cn, kuijia@scut.edu.cn

South China University of Technology

DexForce Co. Ltd.

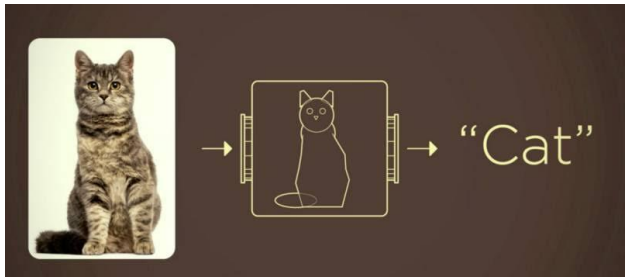
Contents

- **Background**
- **Problem Definition**
- **Related Work**
- **Comprehensive Study**
- **Future Work**

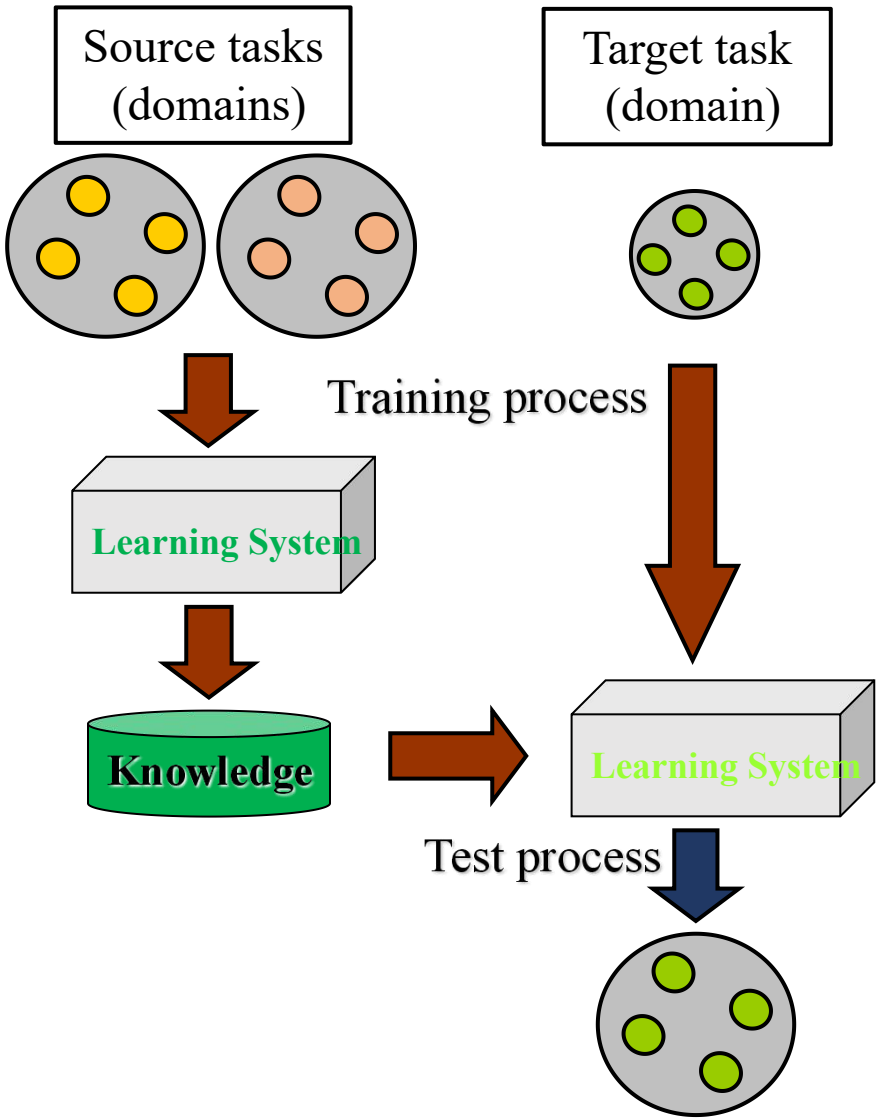
Background

Wow! Deep learning of neuron networks has achieved great success in many computer vision tasks, such as image classification.

But the success relies on a large amount of training data. Collecting and annotating data for all domains and tasks is extremely expensive and time-consuming.

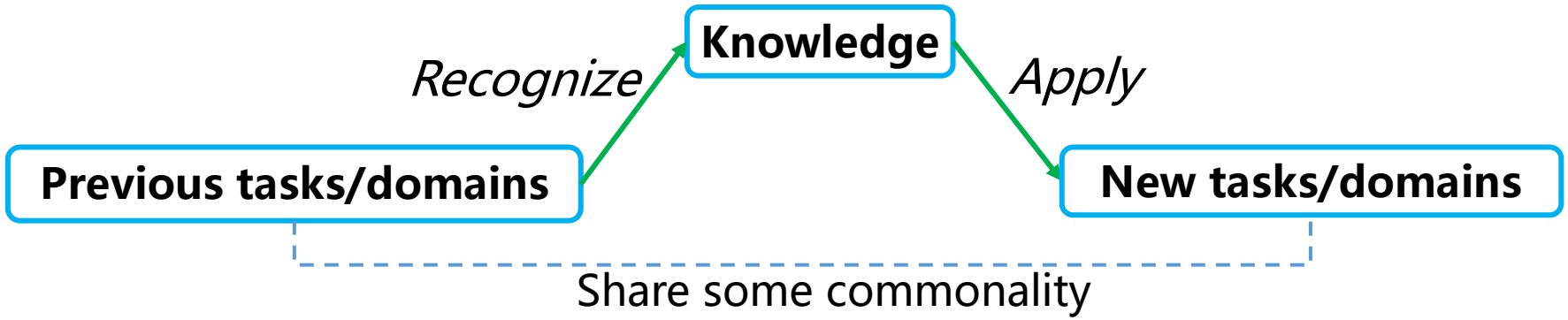


Background



We can tackle it with data-efficient learning, such as **transfer learning, unsupervised domain adaptation**, semi-supervised learning, and few-shot learning.

Problem Definition



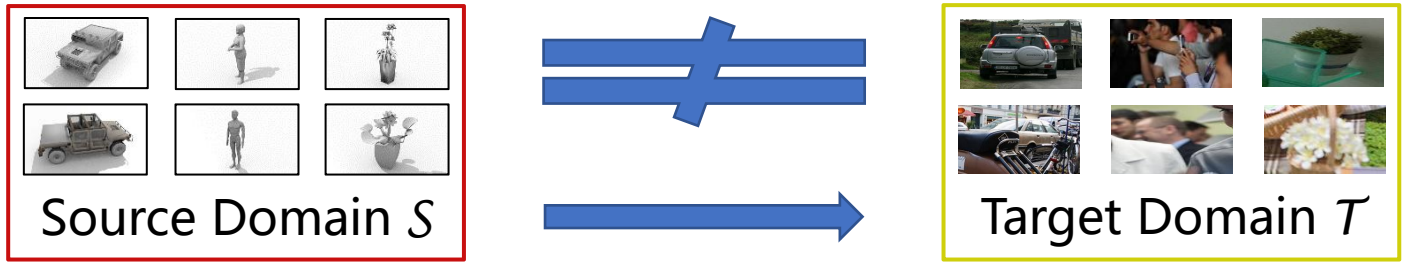
- **Goal: If the new task lacks high-quality training data, the knowledge from the previous task can be transferred to the new task.**
- **Focus: The typical scenario of domain adaptation — transfer knowledge of synthetic data to help classify real data.**

Problem Definition

Unsupervised Domain Adaptation (UDA)

- When the source and target domains have different distributions but share all or part of semantic label space, transfer learning is equivalent to domain adaptation.

- We consider the most practical setting — unsupervised domain adaptation, where **target samples are all unlabeled.**



Unsupervised Domain Adaptation!

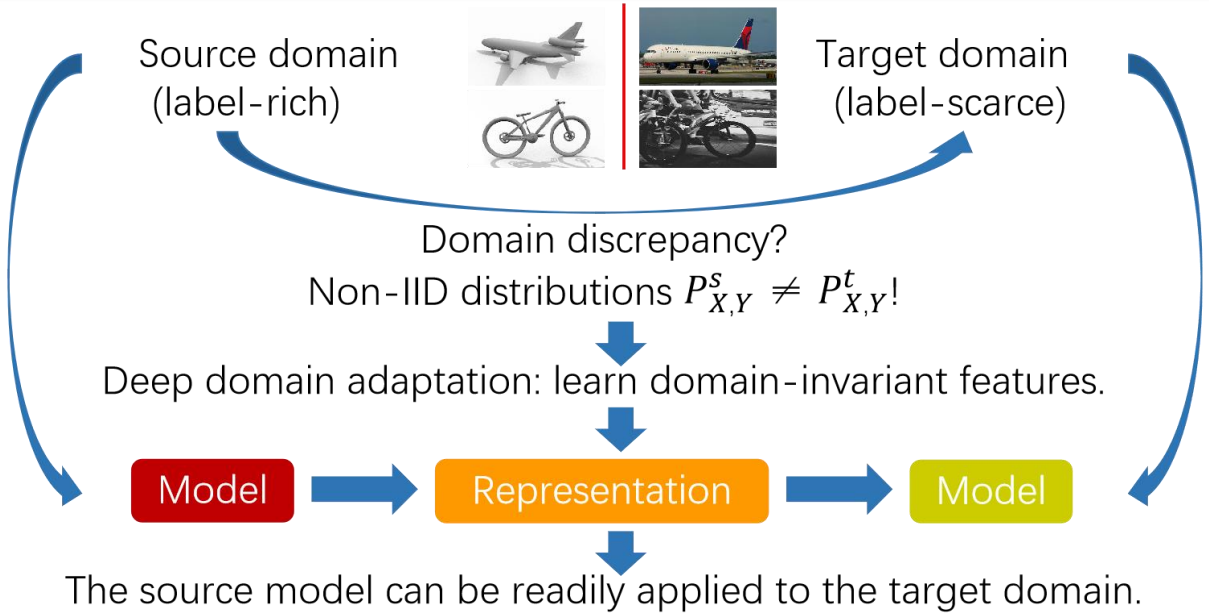
Problem Definition

- Source domain $S = \{(x_j^s, y_j^s)\}_{j=1}^{n_s}$
- Target domain $T = \{x_i^t\}_{i=1}^{n_t}$

A shared label space Y: $y^s, y^t \in \{1,2,\ldots,K\}$

- Feature embedding function $\varphi(\cdot;\theta): X \to Z$ lifts any $x \in X$ to the feature space Z, i.e. $z = \varphi(x)$.
- Classifier $f(\cdot;\vartheta): Z \to R^K$ with softmax at the top outputs a probability vector $p = \text{softmax}(f(z))$.

➢ **Objective: Given labeled data on S, UDA is to predict class labels for unlabeled data sampled from T by learning $\varphi(\cdot)$ and $f(\cdot)$ on both $\{(x_j^s, y_j^s)\}_{j=1}^{n_s}$ and $\{x_i^t\}_{i=1}^{n_t}$.**



Related Work

- **Synthetic Datasets:**
  - VisDA-2017 [1], the first large-scale cross-domain object classification dataset, tailored for domain adaptation from simulation (152K) to reality (55K) across 12 classes.
  - Generated by 3D rendering [2] and domain randomization [3].

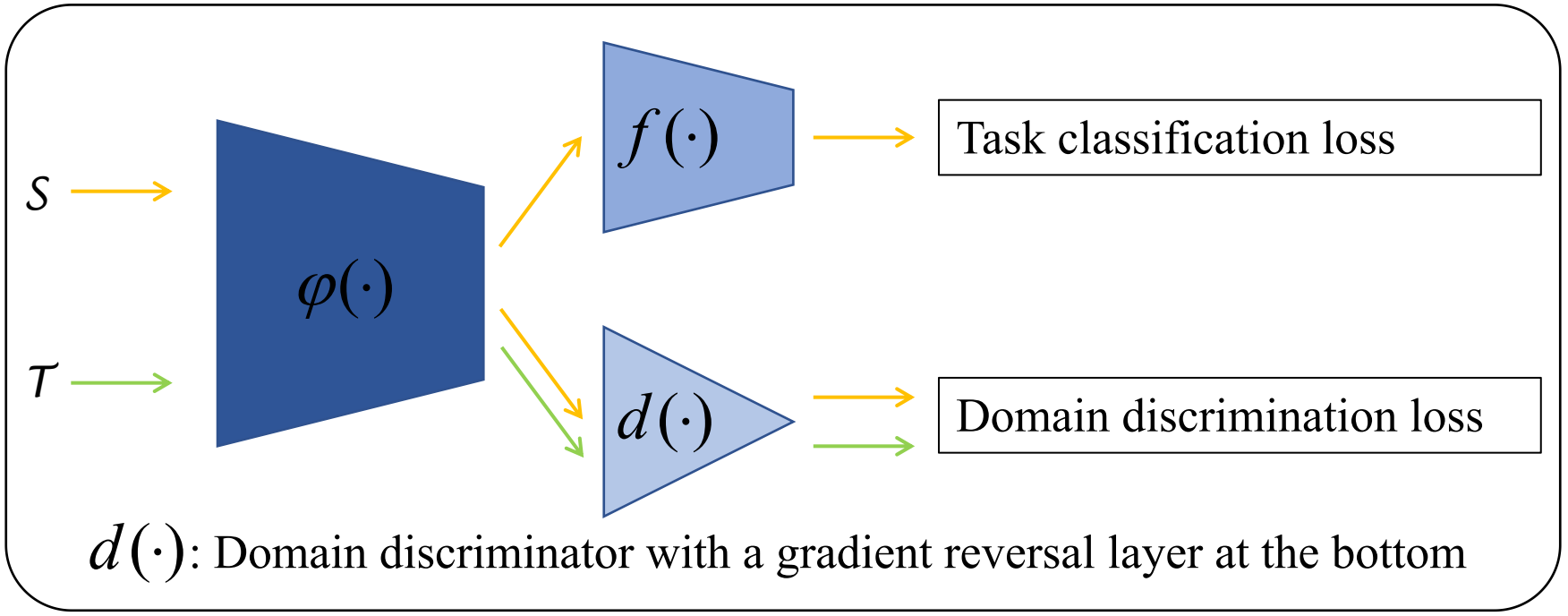
- **Pre-training and Then Fine-Tuning:**
  - Large-scale real data pre-training, e.g., utilizing JFT-300M [4] to study the influence of pre-training on downstream vision tasks.

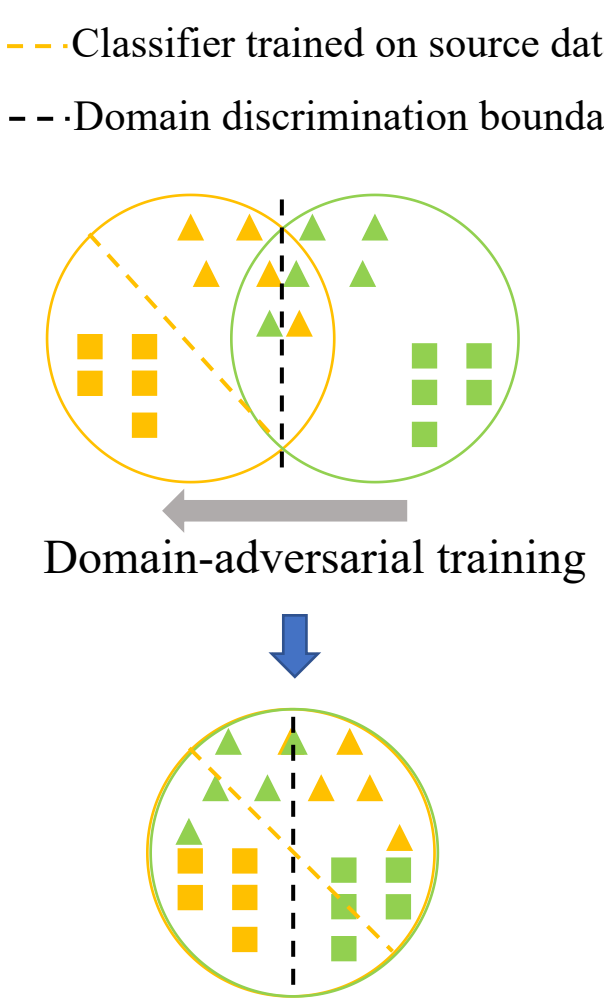
[1] Xingchao Peng, Ben Usman, Neela Kaushik, et al.. VisDA: A Synthetic-to-Real Benchmark for Visual Domain Adaptation. CVPRW, 2018.
[2] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam. Blenderproc. Preprint arXiv:1911.01911, 2019.
[3] J. Tobin et al.. Domain randomization for transferring deep neural networks from simulation to the real world. IEEE International Conference on Intelligent Robots and Systems, 2017.
[4] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. ICCV, 2017.

Related Work

Domain-Level Feature Alignment

- **DANN [5] leverages a domain adversarial task to align the source and target domains as a whole, such that class labels can be transferred from source to target.**



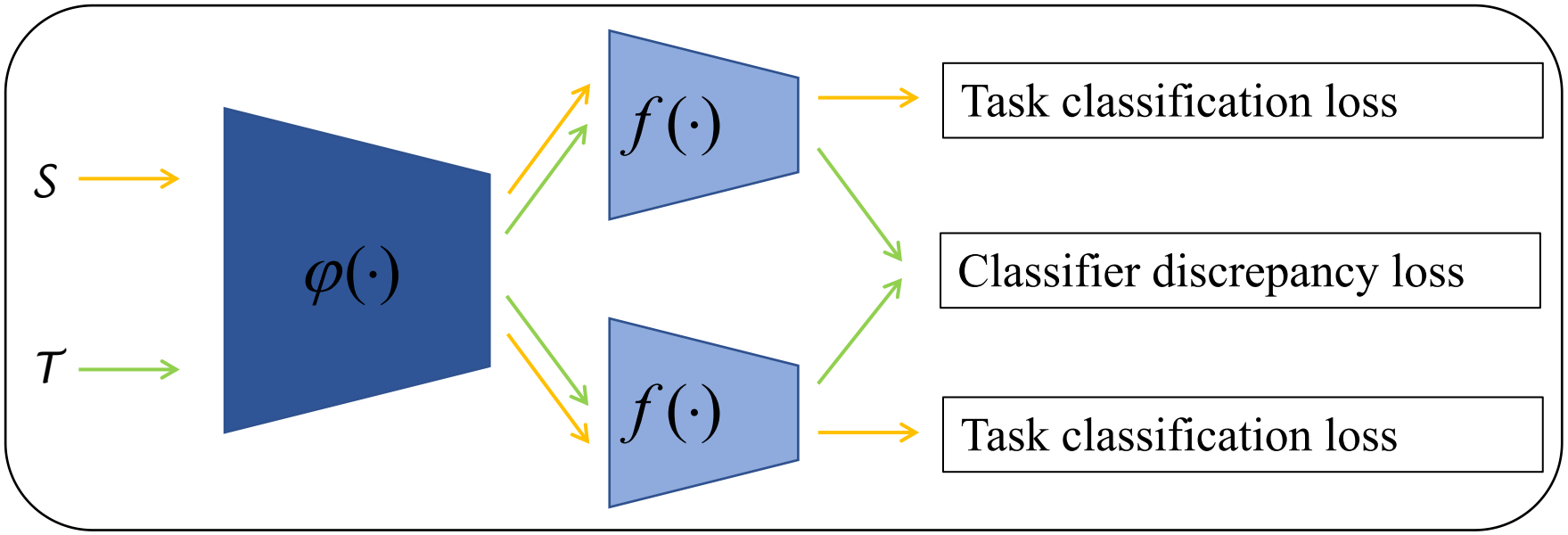


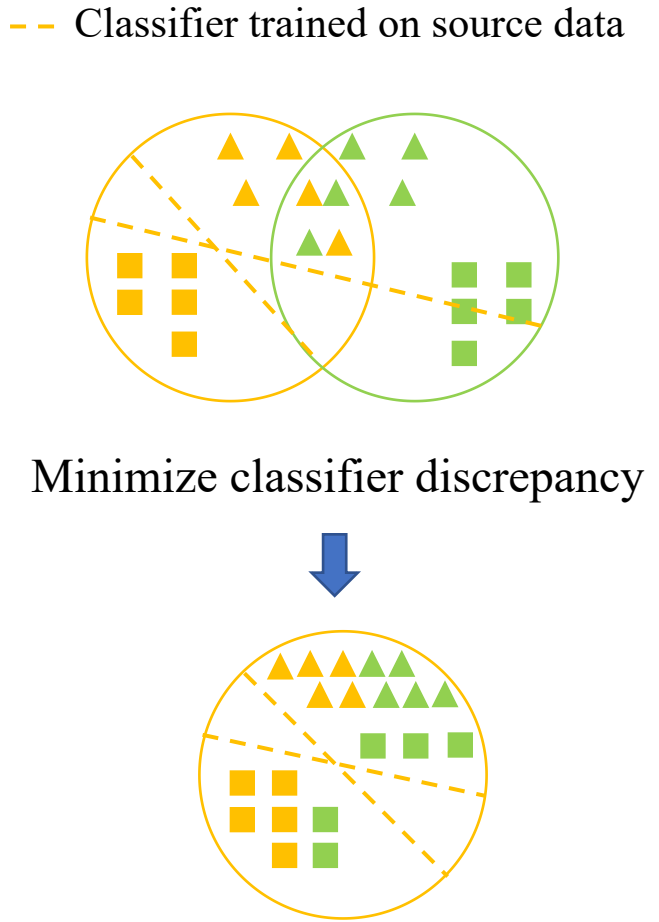
[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, et al.. Domain-adversarial training of neural networks. JMLR, 2016.

Related Work

Class-Level Feature Alignment

- **MCD [6] uses individual task classifiers for the two domains to detect non-discriminative features by maximizing the classifier discrepancy and reversely learn a discriminative feature extractor by minimizing the classifier discrepancy.**





[6] K. Saito, K.Watanabe, Y. Ushiku, et al.. Maximum classifier discrepancy for unsupervised domain adaptation. CVPR, 2018.

Comprehensive Study

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

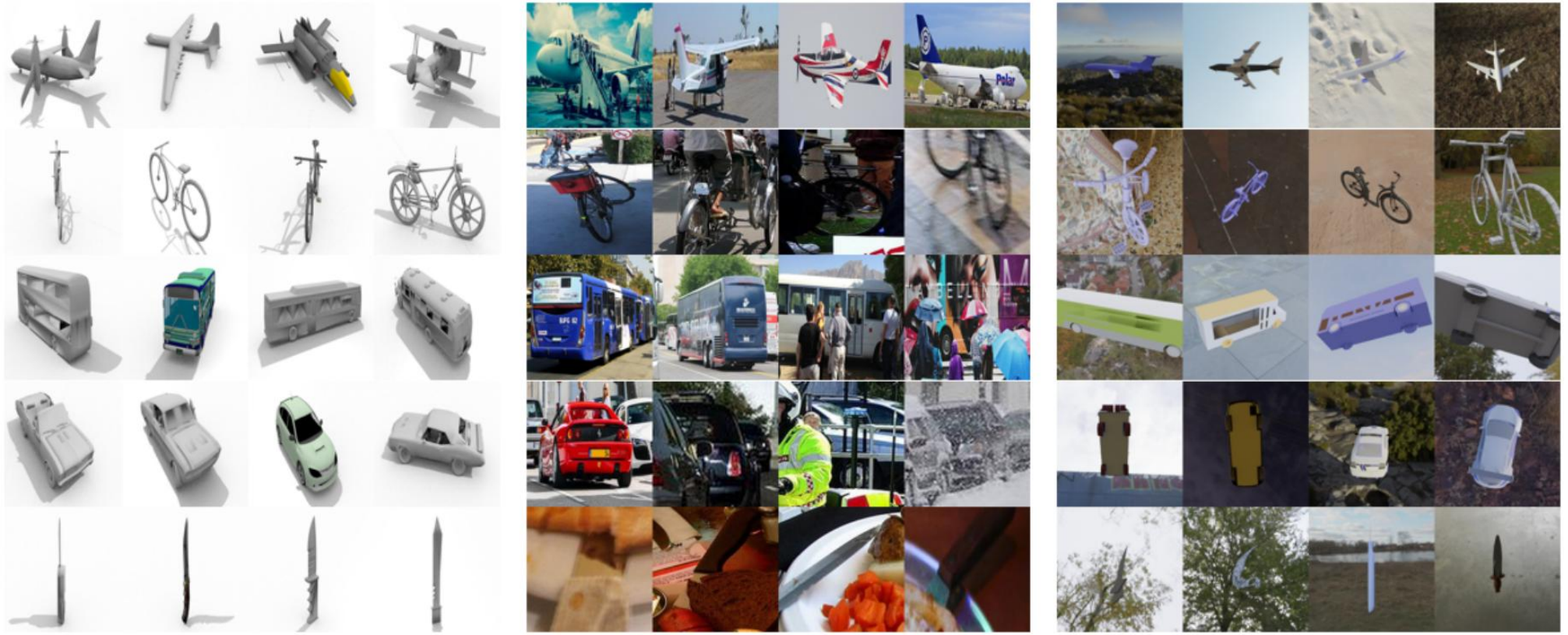
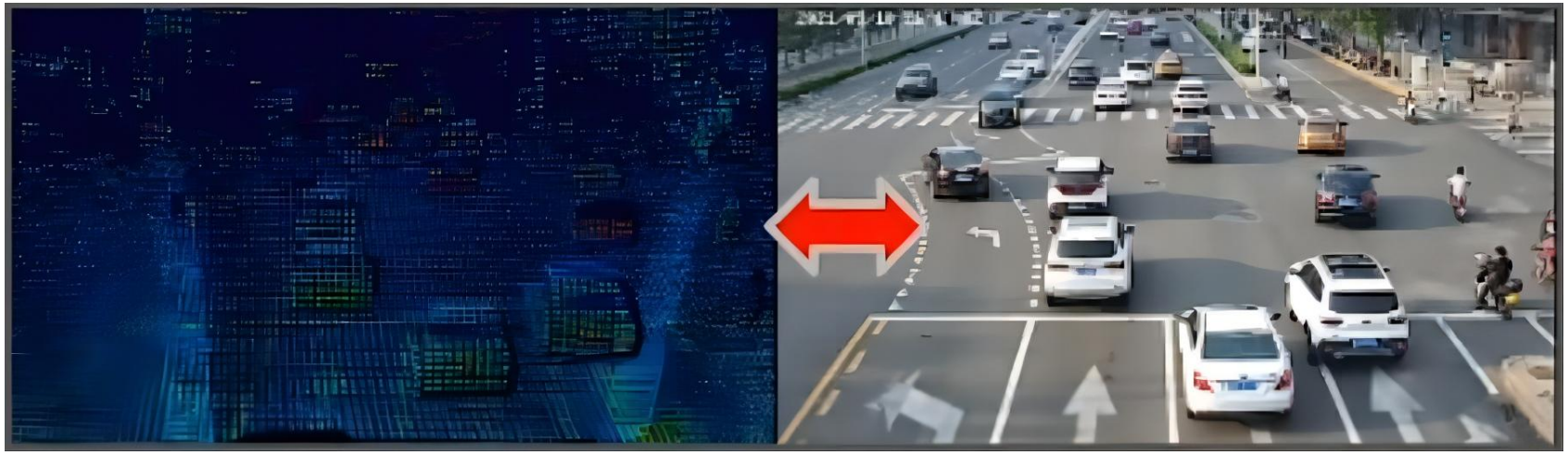


Fig: Sample images from the training (left) and validation (middle) domains of VisDA-2017 and our synthesized data (right).

Comprehensive Study

- Basic and important problems in the context of image classification:
  - Lack of comprehensive synthetic data research.
  - Insufficient exploration of synthetic-to-real transfer.

- Use a 3D rendering engine to build large-scale synthetic datasets (theoretically infinite) and do a comprehensive study on supervised learning and downstream transferring.



[7] 邓志东。AI与自动驾驶：人工智能有可能实现人类智能的挑战性任务吗？AI TIME，2022年09月16日。

Comprehensive Study

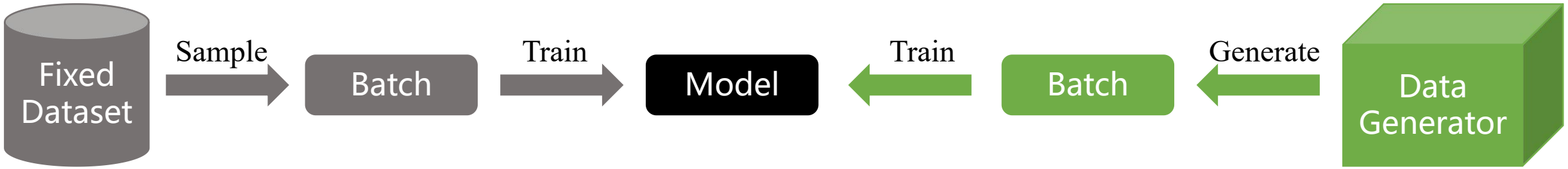
Contributions — explore the answers to the following interesting problems:

- 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?

Comprehensive Study — Bare Supervsied Learning

- Existing works verify classical theories and reveal new findings on real data:
  - The process of acquiring real data cannot be controlled.
  - The annotation accuracy cannot be guaranteed.
  - There may be duplicate images in the training set and test set.
  - The training set and test set are no longer IID.

- Use 3D rendering and domain randomization to generate IID synthetic data:
  - Verify learning insights on shortcut learning, PAC generalization, and variance-bias trade-off.
  - Explore the effects of changing data regimes and network structures on model generalization.
  - **Key design:** the traditional fixed-dataset periodic training vs. a new strategy of training on non-repetitive samples.



Comprehensive Study — Bare Supervsied Learning

- Training data:
  - SubVisDA-10: 10 object classes common in VisDA-2017 [1] and ShapeNet [8], 130, 725 synthetic images, 46, 697 real images.
  - Our 120K synthetic images.
  - Our mutually exclusive batches of synthesized samples per iteration (12.8M in total).

- Test data:
  - IID data: 60K samples that follow the same distribution as our synthesized training data.
  - IID data without background: 60K images to examine the dependency of network predictions on contexts.
  - OOD data: real images from SubVisDA-10.

- Network structures:
  - ResNet-50 [9].
  - ViT-B [10].
  - Mixer-B [11].

[8] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. ArXiv:1512.03012 [cs.GR], Stanford University --- Princeton University --- Toyota Technological Institute at Chicago, 2015.
[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CVPR, 2016.
[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.
[11] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, et al. MLP-Mixer: An all-MLP Architecture for Vision. NeurIPS, 2021.

Comprehensive Study — Bare Supervsied Learning

- **Fixed-Dataset Periodic Training vs. Training on Non-Repetitive Samples:**

  - With strong data augmentation, 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.

Tab: Fixed-dataset periodic training vs. training on non-repetitive samples.
FD: Fixed Dataset, True (T) or False (F). DA: Data Augmentation, None (N), Weak (W), or Strong (S). BG: BackGround.

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**

Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
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

Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
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**

Comprehensive Study — Bare Supervsied Learning

- **Evaluating Various Network Architectures:**

  - In IID tests, ViT performs surprisingly poorly whatever the data augmentation is and even the triple number of training epochs does not improve much.

Tab: Fixed-dataset periodic training vs. training on non-repetitive samples.
FD: Fixed Dataset, True (T) or False (F). DA: Data Augmentation, None (N), Weak (W), or Strong (S). BG: BackGround.

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**

Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
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

Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
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**

Comprehensive Study — Bare Supervsied Learning

- **Impact of Model Capacity & Impact of Training Data Quantity:**

  - There is always a bottleneck from synthetic data to OOD/real data, where increasing data size and model capacity brings no more benefits, and DA to bridge the distribution gap is indispensable except for evolving the image generation pipeline to synthesize more realistic images.

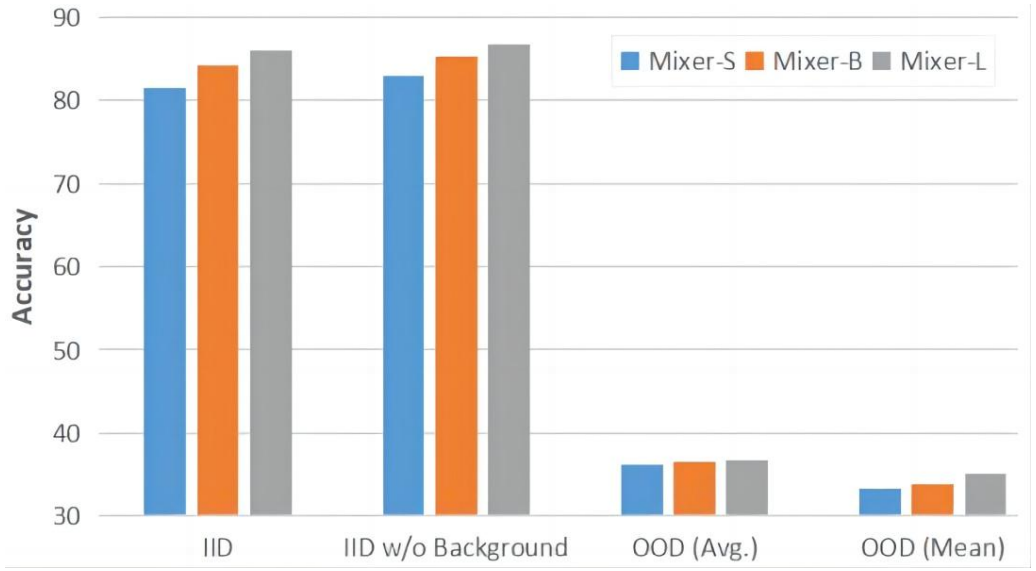


Fig: Generalization accuracy w.r.t. model capacity.

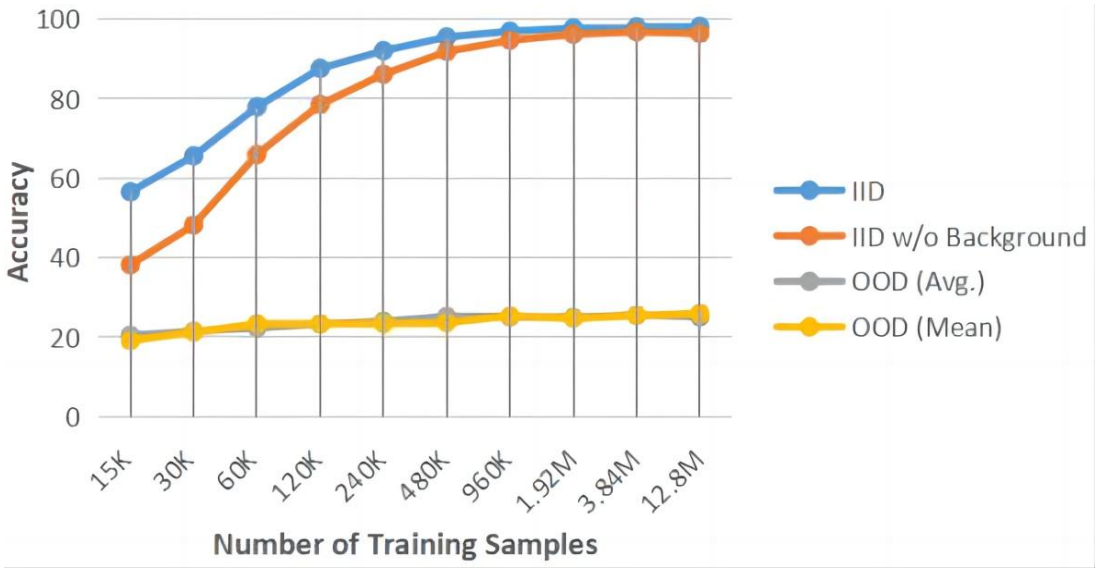


Fig: Generalization accuracy w.r.t. training data quantity.

Comprehensive Study — Bare Supervsied Learning

- **Impact of Data Augmentations:**

  - For the data-unrepeatable training, IID and OOD generalizations are some type of zero-sum game w.r.t. the strength of data augmentation.

Tab: Fixed-dataset periodic training vs. training on non-repetitive samples.
FD: Fixed Dataset, True (T) or False (F). DA: Data Augmentation, None (N), Weak (W), or Strong (S). BG: BackGround.

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**

Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
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

Data	FD	DA	IID Acc./Mean	IID w/o BG Acc./Mean	OOD Acc.	OOD Mean
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**

Comprehensive Study — Assessing Image Variation Factors

- **Explore how variation factors of an image affect the model generalization:**
  - Object scale, material texture, illumination, camera viewpoint, and background.
  - Different rendering variation factors and even their different values have uneven importance to model generalization.
  - Stress the under-explored topic of data generation — AutoSimulate/Weighted Rendering [12].

Tab: Fix vs. randomize image variation factors.

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

[12] H. S. Behl, A. G. Baydin, R. Gal, P. H. S. Torr and V. Vineet. AutoSimulate: (Quickly) Learning Synthetic Data Generation. ECCV, 2020.

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- Bare supervised learning on synthetic data results in poor performance in OOD/real tests:
  - Pre-training and then domain adaptation can improve.
  - However, there is little research exploring the effects of pre-training on DA.

- Study how different pre-training schemes including synthetic data pre-training affect the practical, large-scale synthetic-to-real classification adaptation.

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- Pre-training data:
  - Ours: our synthesized 120K images of the 10 object classes shared by SubVisDA-10.
  - SynSL: our synthesized 12.8M images of the 10 classes for supervised learning.
  - SubImageNet: 25,686 images, the subset collecting examples of the 10 classes from ImageNet [13].
  - Ours+SubImageNet: our synthesized 120K images combined with SubImageNet.
  - ImageNet-990: the fine-grained subclasses for each of the 10 classes are merged into one.
  - ImageNet-990+Ours: ImageNet-990 combined with our 120K synthetic images.
  - ImageNet: the full set of ImageNet (1K classes).
  - MetaShift: 2.56M [14].
- Downstream task: SubVisDA-10.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. Imagenet: A large-scale hierarchical image database. CVPR, 2009.
[14] W. Liang and J. Zou. MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. ICLR, 2022.

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- **The Importance of Pre-training for DA:**
  - DA fails without pre-training.

Tab: Comparing different pre-training schemes.

⋆ : Official checkpoint. Green or red: Best Acc. or Mean in each row. Ours w. SelfSup: Sup. pre-training with contrastive learning.

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

[15] S. Cicek and S. Soatto. Unsupervised domain adaptation via regularized conditional alignment. ICCV, 2019.
[16] H. Tang, K. Chen and K. Jia. Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering. CVPR, 2020.
[17] H. Tang, Y. Wang and K. Jia. Unsupervised domain adaptation via distilled discriminative clustering. Pattern Recognition, 2022.

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- **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.

Tab: Comparing different pre-training schemes.

⋆ : Official checkpoint. Green or red: Best Acc. or Mean in each row. Ours w. SelfSup: Sup. pre-training with contrastive learning.

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

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- **Synthetic Data Pre-training vs. Real Data Pre-training:**

  - Synthetic data pre-training is comparable to or better than real data pre-training — Synthetic data pretraining is promising.

Tab: Comparing different pre-training schemes.

⋆ : Official checkpoint. Green or red: Best Acc. or Mean in each row. Ours w. SelfSup: Sup. pre-training with contrastive learning.

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

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- **Synthetic Data Pre-training vs. Real Data Pre-training:**

  - Synthetic data pre-training is comparable to or better than real data pre-training — Synthetic data pretraining is promising.

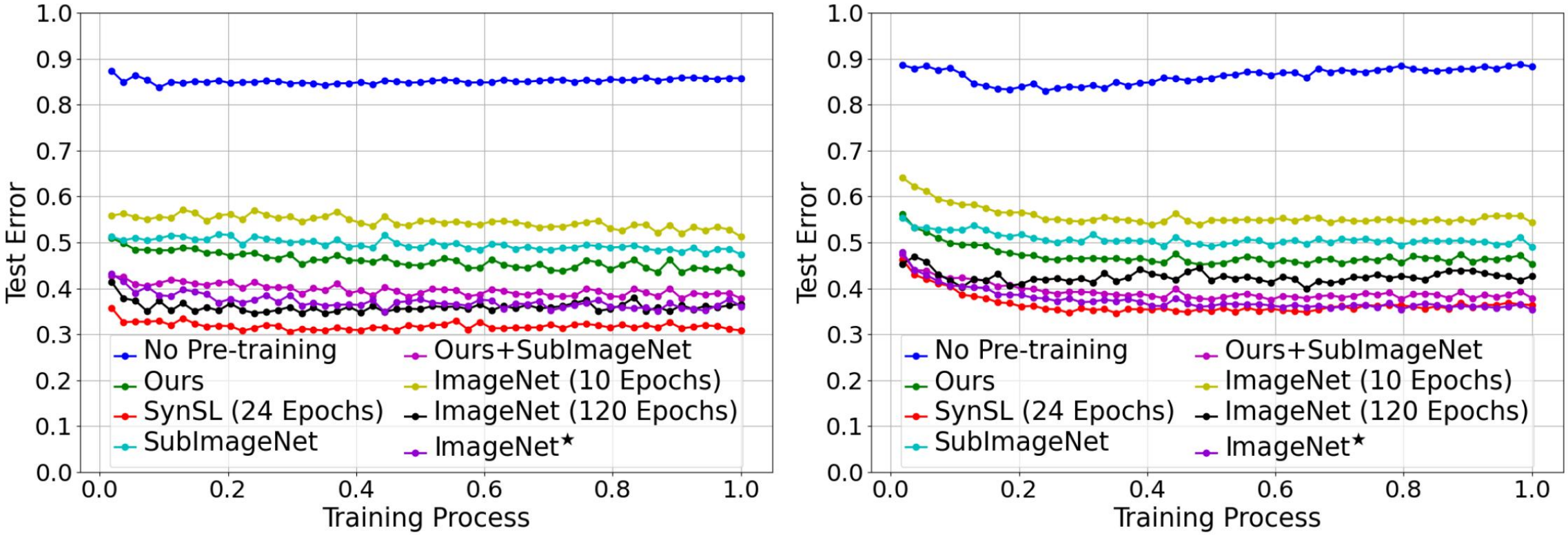


Fig: Learning process (Mean) of MCD (left) and DisClusterDA (right) when varying the pre-training scheme.

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- **Implications for Pre-training Data Setting:**
  - Big Synthesis Small Real is worth deeply researching.
  - Pre-train with target classes first under limited computing resources.

Tab: Comparing different pre-training schemes.

★ : Official checkpoint. Green or red: Best Acc. or Mean in each row. Ours w. SelfSup: Sup. pre-training with contrastive learning.

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

Comprehensive Study — Exploring Pre-training for Domain Adaptation

- **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.

Tab: Comparing different pre-training schemes.

⋆ : Official checkpoint. Green or red: Best Acc. or Mean in each row. Ours w. SelfSup: Sup. pre-training with contrastive learning.

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

Comprehensive Study — A New Synthetic-to-Real Benchmark

- **Introduce a new, large-scale synthetic-to-real benchmark for classification adaptation (S2RDA):**
  - S2RDA-49 + S2RDA-MS-39.
  - Provide a baseline performance analysis for representative DA approaches.
  - Set a more practical and challenging benchmark for future DA research.

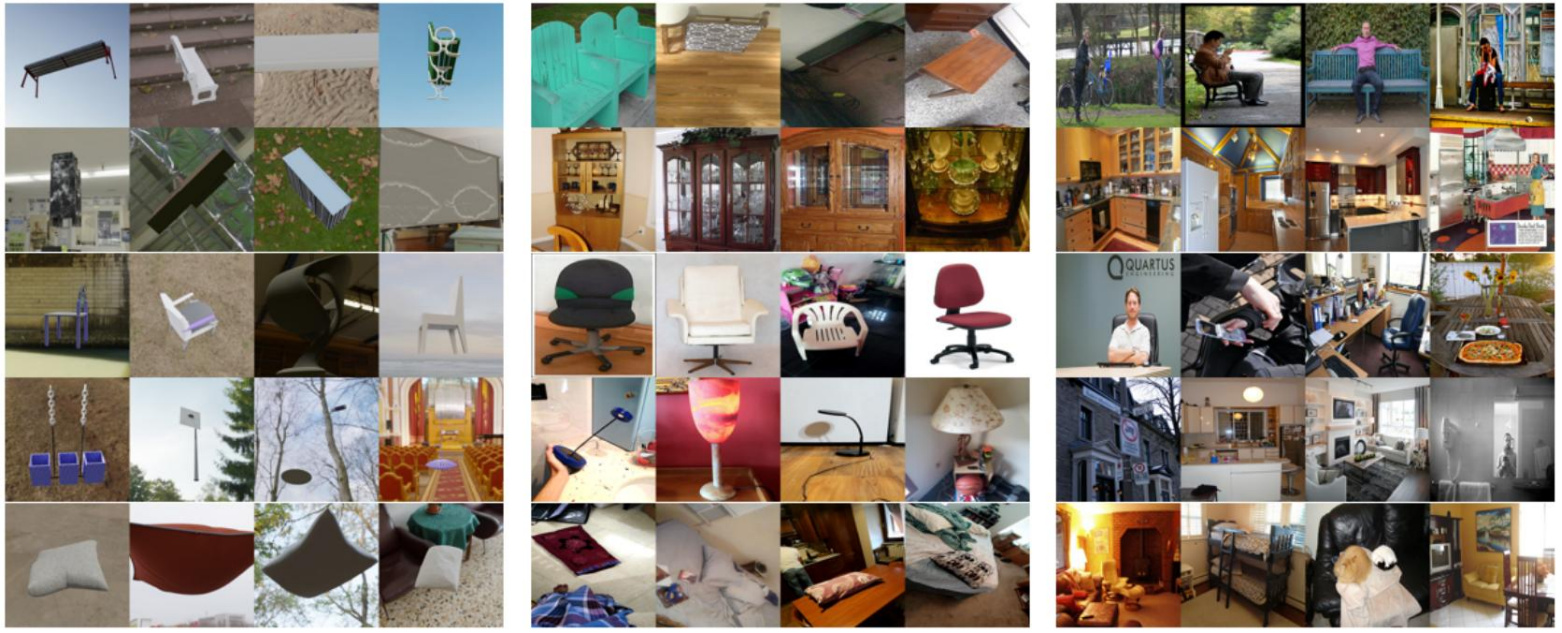


Fig: Sample images from the synthetic (left) domain and the real domains of S2RDA-49 (middle) and S2RDA-MS-39 (right).

Tab: Domain adaptation performance on S2RDA.

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**

Future Work

01 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.

02 Evaluation metrics robust to pre-training

The comparison among various DA methods yields different or even opposite results when using different pre-training schemes. DA researchers should propose and follow evaluation metrics enabling effective and fair comparison.

03 More realistic simulation synthesis

We will consider more imaging parameters, e.g., randomizing the type and hue of the light, including physical objects with actual textures from YCB, and using the flying distractor.

04 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 is to be explored.

05 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 and HyperSim for segmentation and object detection.

A Takeaway Message

To solve the basic and important problems in the context of image classification, such as the lack of comprehensive synthetic data research and the insufficient exploration of synthetic-to-real transfer, we propose to exploit synthetic datasets to explore questions on model generalization, benchmark pre-training strategies for DA, and build a large-scale benchmark dataset S2RDA for synthetic-to-real transfer, which can push forward future DA research.

Code and datasets

References

[1] Xingchao Peng, Ben Usman, Neela Kaushik, et al.. VisDA: A Synthetic-to-Real Benchmark for Visual Domain Adaptation. CVPRW, 2018.

[2] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam. Blenderproc. Preprint arXiv:1911.01911, 2019.

[3] J. Tobin et al.. Domain randomization for transferring deep neural networks from simulation to the real world. IEEE International Conference on Intelligent Robots and Systems, 2017.

[4] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. ICCV, 2017.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, et al.. Domain-adversarial training of neural networks. JMLR, 2016.

[6] K. Saito, K.Watanabe, Y. Ushiku, et al.. Maximum classifier discrepancy for unsupervised domain adaptation. CVPR, 2018.

[7] 邓志东。AI与自动驾驶：人工智能有可能实现人类智能的挑战性任务吗？AI TIME，2022年09月16日。

[8] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. ArXiv:1512.03012 [cs.GR], Stanford University --- Princeton University --- Toyota Technological Institute at Chicago, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CVPR, 2016.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

[11] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, et al. MLP-Mixer: An all-MLP Architecture for Vision. NeurIPS, 2021.

[12] H. S. Behl, A. G. Baydin, R. Gal, P. H. S. Torr and V. Vineet. AutoSimulate: (Quickly) Learning Synthetic Data Generation. ECCV, 2020.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. Imagenet: A large-scale hierarchical image database. CVPR, 2009.

[14] W. Liang and J. Zou. MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. ICLR, 2022.

[15] S. Cicek and S. Soatto. Unsupervised domain adaptation via regularized conditional alignment. ICCV, 2019.

[16] H. Tang, K. Chen and K. Jia. Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering. CVPR, 2020.

[17] H. Tang, Y. Wang and K. Jia. Unsupervised domain adaptation via distilled discriminative clustering. Pattern Recognition, 2022.


Thanks for listening !