© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Luo, Yiran, Joshua Feinglass, Tejas Gokhale, Kuan-Cheng Lee, Chitta Baral, and Yezhou Yang. "Grounding Stylistic Domain Generalization with Quantitative Domain Shift Measures and Synthetic Scene Images." 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2024, 7303–13. https://doi.org/10.1109/CVPRW63382.2024.00726.

https://doi.org/10.1109/CVPRW63382.2024.00726

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

arXiv:2405.15961v1 [cs.CV] 24 May 2024

Grounding Stylistic Domain Generalization with Quantitative Domain Shift Measures and Synthetic Scene Images

Yiran Luo[1*] Joshua Feinglass[1] Tejas Gokhale[2] Kuan-Cheng Lee[1] Chitta Baral[1] Yezhou Yang[1]
[1]Arizona State University [2]University of Maryland, Baltimore County
[1]{yluo97, jfeingla, klee201, chitta, yz.yang}@asu.edu [2]gokhale@umbc.edu

Abstract

*Domain Generalization (DG) is a challenging task in machine learning that requires a coherent ability to comprehend shifts across various domains through extraction of domain-invariant features. DG performance is typically evaluated by performing image classification in domains of various image styles. However, current methodology lacks quantitative understanding about shifts in stylistic domain, and relies on a vast amount of pre-training data, such as ImageNet1K, which are predominantly in photo-realistic style with weakly supervised class labels. Such a data-driven practice could potentially result in spurious correlation and inflated performance on DG benchmarks. In this paper, we introduce a new 3-part DG paradigm to address these risks. We first introduce two new quantitative measures **ICV** and **IDD** to describe domain shifts in terms of consistency of classes within one domain and similarity between two stylistic domains. We then present **SuperMarioDomains (SMD)**, a novel synthetic multi-domain dataset sampled from video game scenes with more consistent classes and sufficient dissimilarity compared to ImageNet1K. We demonstrate our DG method **SMOS**. SMOS uses SMD to first train a precursor model, which is then used to ground the training on a DG benchmark. We observe that SMOS+SMD altogether contributes to state-of-the-art performance across five DG benchmarks, gaining large improvements to performances on abstract domains along with on-par or slight improvements to those on photo-realistic domains. Our qualitative analysis suggests that these improvements can be attributed to reduced distributional divergence between originally distant domains. Our data are available at https://github.com/fpsluozi/SMD-SMOS .*

1. Introduction

The generalizability of deep neural networks in computer vision is a crucial task that is still challenging [3, 18].

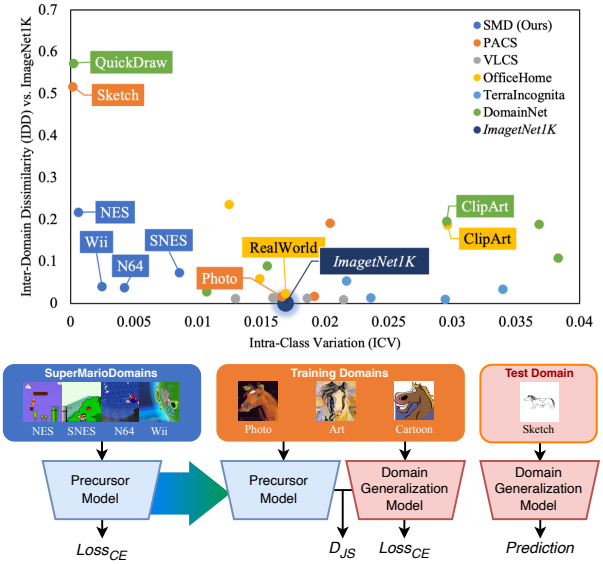


Figure 1. **Top:** We define two quantitative measures **ICV** and **IDD** to describe stylistic domain shifts in image datasets for Domain Generalization (DG). We find that the vast ImageNet1K, commonly used for pre-training DG models, has inconsistent class labels and is already similar in style with photo-realistic domains found in multiple benchmarks. Therefore, we compile a novel synthetic dataset **SuperMarioDomains (SMD)** as referential stylistic domains with consistent scene class labels and sufficient dissimilarity from existing domains. **Bottom:** We present our DG approach **SMOS** that leverages the unique domain shifts in our new **SMD** dataset. We first train a Precursor Model using SMD and cross entropy $Loss_{\text{CE}}$. We then utilize the trained Precursor Model to ground the training of the DG model with training domains from the benchmark, optimizing the empirical loss of both cross entropy $Loss_{\text{CE}}$ and Jensen-Shannon Divergence D_{JS} between the Precursor Model and the DG Model.

Hence, the specific task of Domain Generalization (DG) has been defined with the aim of improving the generalization of models by singling out distribution shifts among data that belong to independent and identically distributed (i.i.d) domains [22, 44]. The evaluation of a DG model consists

*Corresponding author

of performing supervised image classification in a multi-source leave-one-out scenario, where one domain is held out as an unseen test set, and other domains are for training.

The crucial DG strategy is to learn domain-invariant features from the training domains such as SagNet [42], CORAL [53], and DANN [17]. More DG benchmarks of more perplexing stylistic domains and more fine-grained classes [4, 5, 31, 46, 57] have also been developed for more comprehensive evaluation of generalizability. However, we notice potential risks in the current methodology. Most DG methods solely rely on initializing their backbone models pre-trained with vast weakly supervised image collections, e.g. ImageNet1K [13], which overwhelmingly resemble [38, 55] photo-styled domains in multiple DG benchmarks. We observe that many DG methods [2, 9, 42, 53, 56] reduce various forms of qualitative distances among domains, but lack quantitative understanding of the specific differences amongst the domains of image styles. Without clarified and unbiased understanding of the common ground among training domain examples, generalization onto an unseen domain would potentially be based upon un-grounded spurious evidence, resulting in inflated DG performance.

In this paper, we present a new 3-fold paradigm for Domain Generalization. First, we define 2 quantitative measures for stylistic domain shifts based on Jensen-Shannon Divergence [40] - **Intra-Class Variation (ICV)** within an individual domain, and **Inter-Domain Dissimilarity (IDD)** between two domains. With these 2 measures, as shown in Figure 1 in the top, we confirm that ImageNet1K is biased toward domains of photorealistic styles, but also has inconsistent representations within its individual class categories.

We then construct a new multidomain dataset dubbed **SuperMarioDomains (SMD)** [1]. SMD features 4 domains representing multiple generations of video game graphic styles, ranging from low-resolution pixels to 3D-rendered polygon-rich graphics. All domains share 4 classes of in-game type of scenes that appear consistently throughout the Mario franchise. Unlike ImageNet1K, SMD's domains maintain variable stylistic distances from ImageNet1K in terms of IDD, while having more consistently labeled examples than ImageNet1K's in terms of ICV.

Our proposed DG method **SMOS** is shown in the bottom of Figure 1. SMOS first trains a precursor model with domains in SMD. It then trains a DG model grounded by the distribution represented by the SMD-trained precursor model. We apply SMOS and find overall improvements on multiple DG benchmarks. SMOS contributes to the most significant improvements when targeting abstract-styled domains, for example $+7.3\%$ on Sketch of PACS [36], $+3.6\%$ on Clipart of OfficeHome [57], or $+8.1\%$ on Clipart of DomainNet [57] over the baselines of MIRO [9]. SMOS also maintains performance on par with baselines when targeting photo-realistic domains, and in some cases, e.g. PACS, even has improvements. We also observe that SMOS is able to improve extraction of domain-invariant features and generalization over domain shifts in quality, as we find that originally distant stylistic domains are now projected within considerably smaller distributional divergence.

Dataset	Domains	Classes	Size
ImageNet1K [13]	1 image style	1K objects	1.3M
Places365 [65]	1 image style	365 scenes	1.8M
PACS [36]	4 image styles	7 objects	10K
VLCS [29]	4 photo sources	5 objects	10K
TerraIncognita [5]	4 locations	10 animals	25K
OfficeHome [57]	4 image styles	65 objects	16K
DomainNet [46]	6 image styles	345 objects	587K
SMD (ours)	4 game graphics	4 scenes	82K

Table 1. Statistics of image classification datasets involved in DG.

Our main contributions in this paper are as follows.

1. We propose ICV and IDD as measures for stylistic domain shifts in DG benchmarks. Using our measures, we find that real-world datasets like ImageNet1K used as pre-training data for DG may not be ideal for obtaining domain-invariant features due to inconsistent classes and overt stylistic similarity with training domains.
2. We introduce a new synthetic dataset SuperMarioDomains as a precursor dataset for DG, incorporating unique features of consistent classes of video game scenes across stylistic domains in video game graphics that are dissimilar to ImageNet1K.
3. We present our DG method SMOS. SMOS utilizes SuperMarioDomains to obtain a precursor model that grounds the training process on DG benchmarks. We show that SMOS is capable of obtaining top performance on multiple DG benchmarks, in particular, via large improvements on targeting abstract-styled domains together with on-par or slight improvements on photorealistic-styled domains.

2. Related Works

Synthetic Datasets with Domain Shifts. Synthetic data has long been used in various disciplines in computer vision [1, 7, 11, 45, 54, 60]. Recently, we have seen that more synthetic datasets of distribution shifting domains are being assembled to encourage more robust algorithms in different

[1] As of Mar. 20, 2024, the official content guidelines of Nintendo, who owns the copyright of all games involved in SMD, has clearly stated they 'will not object to your use of gameplay footage and/or screenshots captured from games for which Nintendo owns the copyright ... for appropriate video and image sharing sites' [41]. We will follow this guideline and publish our work in two forms. Extracted feature vectors and pre-trained models will be readily available under the MIT license. The SMD dataset with its raw images will be accessible after agreeing to using the dataset for fair-use research purposes only under the Open Database License (ODbL).

tasks. We have benchmarks studying adaptation between synthesized and real-life objects in Syn2Real [47] or SYNTHIA [50]. Super-CLEVR [39] studies more robust visual reasoning skills by constructing domains w.r.t. visual question answering settings. More specifically, we draw great inspiration from the GTAV-Cityscapes challenge [10, 49], adapting from a vast collection of video game landscapes to real-world scenarios for segmentation. All these works show that synthetic datasets, though lacking full realism, may help provide great insight into domain shifts.

Domain Generalization Benchmarks. Early DG benchmarks such as Office [51] or VLCS [29] focus solely on photo-realistic images. Since the time when single-style bias was first exposed by DeCAF [14], more fine-grained image styles have been introduced to the mix of image domains, and we have seen a steady increase in scale for DG datasets, including OfficeHome [57], PACS [36], TerraIncognita [5], SVIRO [12], WILDS [31], Camelyon17 [4], and NICO++ [63]. Currently, the DomainNet dataset [46] is the largest with 587K images, featuring 6 stylistic domains and 345 object categories.

Domain Generalization Approaches. Techniques to tackle the problem of domain shift in Domain Generalization have been rapidly developed over the years. Researchers are no longer restricted to straightforward approaches such as finding linear [26] representations with techniques like interpolation [16, 20, 59, 61] or nonlinear [15] representations in-between domains. In many circumstances, simple methods, such as variants of empirical risk minimization [2, 23], can produce high performance on popular domain generalization benchmarks. But recent approaches to domain generalization can involve adaptation and combination of deep neural network models such as DANN [17] and CDANN [37], leveraging Meta-Learning [34] or Adversarial-Learning [19, 48, 58] to transfer model parameters, using an ensemble of different architectures [38], training strategies [28], or regularization methods to improve generalizability like SWAD [8] or MIRO [9].

3. Preliminaries

In this paper, we focus on the Domain Generalization (DG) performance of image classification by conducting Multi-Source Domain Generalization (MSDG), where we evaluate a model's performance with *leave-one-out cross-validation*. In formal terms, a DG dataset $\mathcal{D}$ is divided into N domains $\{d_1, ...d_N\}$. Each individual domain d_i contains images X^{d_i} paired with their class labels Y^{d_i}. All domains share the same set of image labels. We train a model $\mathcal{M}$ with the training set (training domains) $\mathcal{D}_{tr}$ that uses all data from all-but-one domains $\mathcal{D}_{tr} = \{(X^{d_1}, Y^{d_1}), ...(X^{d_{N-1}}, Y^{d_{N-1}})\}$. The trained model is then evaluated using unseen data from the one held-out test (target) domain as the test set $\mathcal{D}_{te} = \{(X^{d_N}, Y^{d_N})\}$. Noticeably, unlike Domain Adaptation [68], the one test domain $\mathcal{D}_{te}$ in DG is not involved in any training and does not contain samples that overlap with the training domains.

A typical DG model $\mathcal{M} = f \circ g$ consists of two core components: a feature extractor (featurizer) f as the backbone network, followed by a single-layered linear classifier g. Although more complicated methods have been developed over the years, the fundamental approach of **Empirical Risk Minimization** (**ERM**) [56] remains relevant. In the context of DG, the goal of ERM is simply minimizing the classification loss averaged over the $N-1$ training domains given training data $x_i, y_i \in \mathcal{D}_{tr}$:

$$\mathcal{L}_{\text{ERM}} = \frac{1}{N-1} \sum_{x_i, y_i \in \mathcal{D}_{tr}} \mathcal{L}_{\text{CE}}(f \circ g(x_i), y_i). \quad (1)$$

Although straightforward, the latest baseline ERM [23], which applies cross-entropy as the training loss $\mathcal{L}_{\text{CE}}$, is shown to rank high in many DG benchmarks [22, 33, 55].

4. Analyses on Domain Shift of Stylistic Domains and Pre-training Data

Risk in Pre-training Data Used for DG. Most DG methods do not train their model $\mathcal{M}$ entirely from scratch. The model's initial weights are commonly transferred from an existing model, typically pre-trained on ImageNet1K [13]. Researchers [55, 63] have observed that there is an apparent stylistic resemblance between ImageNet1K and many domains from various DG benchmarks, such as the Photo domain in PACS[36], the RealWorld domain in OfficeHome[57], or even all 4 domains in VLCS [29].

Such a common practice of only depending on a vast real-world image dataset can be risky. Qualitatively, the images used for pre-training lack shift in style to adapt from, e.g. ImageNet1K are dominated by real-world photos. In addition, the images in ImageNet1K may be inconsistent within the same class; e.g., images of a class of fish may or may not include a person holding it. Hence, when we perform on DG benchmarks by using a model pre-trained only by such data, we may find it difficult to learn coherent style-invariant presentations in order to generalize over distributional shifts among domains, especially when we have to train with the predominating photo-realistic domains and test on an unseen abstract-styled domain.

Quantitative Measures of Stylistic Domain Shift. We would next like to compare the advantages of different DG pre-training data like ImageNet1K. Inspired by previous works which examine image similarity [21, 64], we define two intuitive measures for domain shift in DG: **Intra-Class Variation (ICV)** to implicate the class-wise presentation inconsistency within one specific domain, and **Inter-Domain Dissimilarity (IDD)** to implicate the distance in representation distributions across two domains of styles.

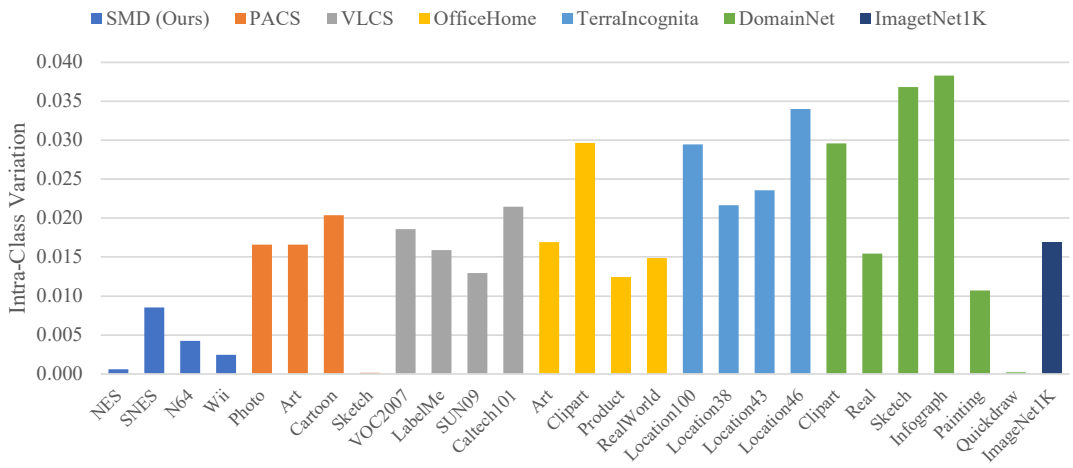


Figure 2. **Intra-Class Variation (ICV)** for each domain in featured datasets. A low ICV indicates that the classes are more consistent in terms of colors, as in NES of SMD, Sketch of PACS, and Quickdraw of DomainNet. Meanwhile, the classes in ImageNet1K, which is commonly used for pre-training in DG, are implicated to be as inconsistent as those in photo- or art-styled domains, e.g. Photo and Art of PACS, LabelMe of VLCS, Art and RealWorld of OfficeHome, as well as Real of DomainNet. Average of 3 trials.

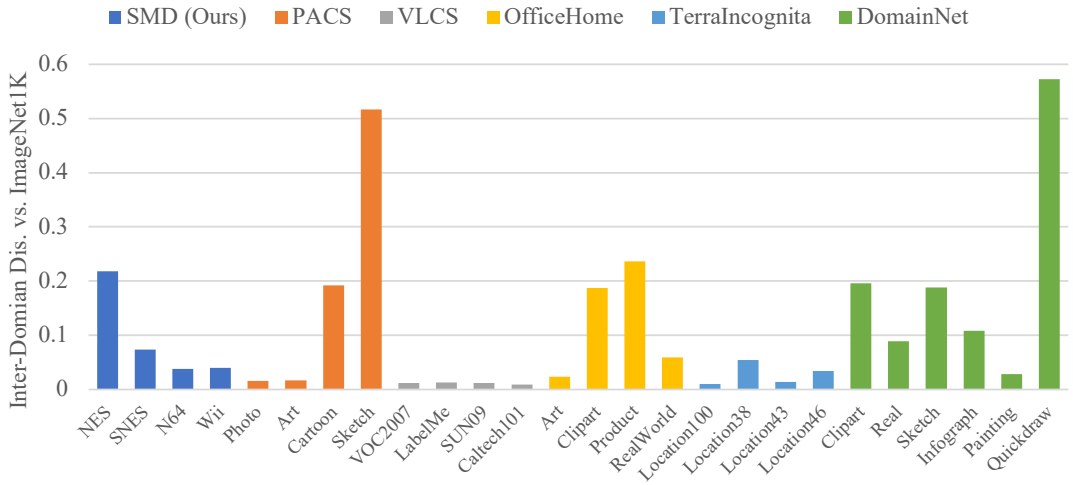


Figure 3. **Inter-Domain Dissimilarity (IDD)** of ImageNet1K vs. each domain in featured datasets. IDD of ImageNet1K vs. itself is presumably 0. Since ImageNet1K is dominantly real-world photographs, those with smaller IDD implicate stronger resemblance to the style of photo-realism, e.g. all 4 domains of VLCS, or Photo of PACS. Highly abstract and simplistic styles, such as Sketch of PACS and Quickdraw of DomainNet, are shown in very large IDD values on the flip side.

Both measures utilize the symmetric Jensen-Shannon Divergence (JSD) [40]. Concretely, let P and Q be the estimated probability distributions of the RGB channels and bins (3×256 dimensions), making the outcome set a vector of length 768. The JSD between P and Q is defined as:

$$D_{\mathrm{JS}}(P,Q) = \frac{1}{2}\left(D_{\mathrm{KL}}(P||M) + D_{\mathrm{KL}}(Q||M)\right), \quad (2)$$

where M is the average mixed distribution:

$$M = \frac{1}{2}(P+Q), \quad (3)$$

and the KL divergence between two distributions P and Q:

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}, \quad (4)$$

where $P(x)$ and $Q(x)$ are the probabilities of image instance x in distributions P and Q respectively. D_{JS} ranges between 0 and 1, with 0 indicating identical distributions and 1 indicating completely dissimilar distributions.

For **Intra-Class Variation** of a given image domain, P_i and Q_i represent distributions of two equal splits of samples

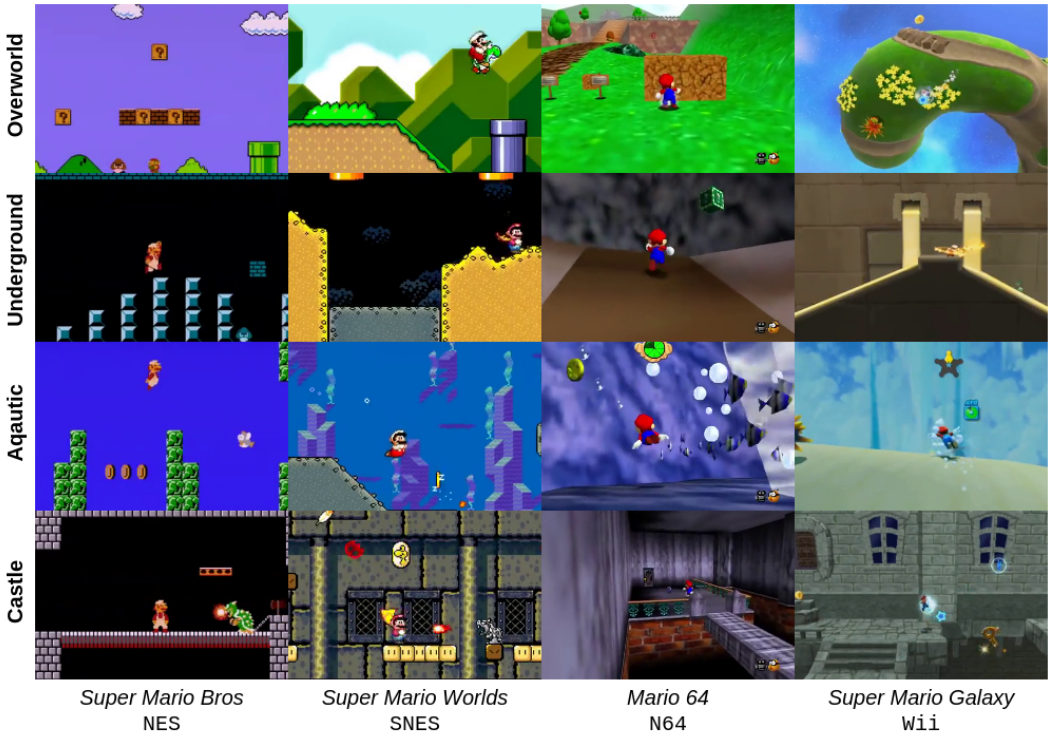


Figure 4. A qualitative overview of our **SuperMarioDomains(SMD)** dataset, consisting of video frames from actual game footage categorized into 4 distinctive scene classes and 4 image style domains. **Columns from left to right:** The four image domains, named after the console hardware on which each game runs - `NES`, `SNES`, `N64`, and `Wii`. **Rows from top to bottom:** The four classes of in-game scenes - Overworld, Underground, Aquatic, and Castle. These synthetic image styles of SMD are drastically different from those in existing DG benchmarks, such as realistic photographs, pencil sketches, or oil paintings.

belonging to the same class i. The **ICV** w.r.t. the domain $\mathcal{D}$ of n classes is the average over all its class-wise JSDs:

$$ICV(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} D_{\text{JS}}(P_i, Q_i). \tag{5}$$

For **Inter-Domain Dissimilarity** between two domains $\mathcal{D}_1$ and $\mathcal{D}_2$, the metric straightforwardly computes the JSD between respective distributions $P_{\mathcal{D}_1}$ and $P_{\mathcal{D}_2}$:

$$IDD(\mathcal{D}_1, \mathcal{D}_2) = D_{\text{JS}}(P_{\mathcal{D}_1}, P_{\mathcal{D}_2}). \tag{6}$$

For fairness, we assume that all image presentations follow Gaussian distributions with regard to probabilities of raw RGB values within bins of $[0, 255]$, normalized by the uniform mean $[0.5, 0.5, 0.5]$ and standard deviation $[0.5, 0.5, 0.5]$. In addition to ImageNet1K, we choose to study the domains in the following DG benchmarks: PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet. The ICV for each domain studied is averaged over 3 trials.

In Figure 2, we first compare the ICV of ImageNet1K with that of every domain in the chosen DG benchmarks. We discover that, in ImagetNet1K, image representations within the same class can be as diversely distributed as those learning samples from photo- or art-styled domains in DG benchmarks. Meanwhile, domains of highly abstract styles, such as Sketch in PACS and Quickdraw in DomainNet, have drastically low ICV values compared to other domains in their respective benchmarks. According to Table 1, since ImageNet1K also has at least twice as many classes compared to DG domains, generalization with known samples solely from high-ICV classes would be difficult to learn domain-invariant features, especially when the test domain samples have vastly low variation in feature space.

From Figure 3, we now have a clear view of ImageNet1K's similarity to existing styles in multiple DG benchmarks. This can be observed, for example, from the next-to-0 IDD values of all four domains of VLCS, the Photo and Art domains of PACS, or the Location38 domain of TerraIncognita. In contrast, styles of abstract color patterns or textures, such as the Sketch domain of PACS and the Quickdraw domain of DomainNet, have large IDD values which suggests large distances from ImageNet1K.

5. Methodology

The SuperMarioDomains Dataset. Taking inspirations from previous efforts to construct synthetic datasets with

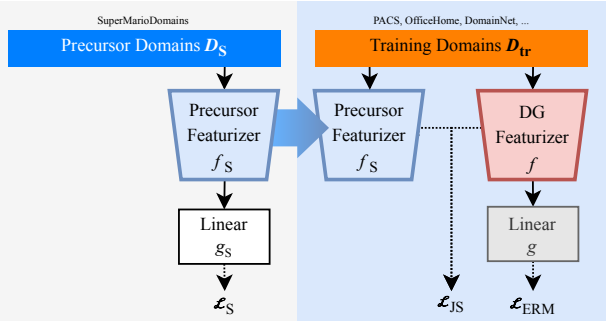


Figure 5. The pipeline of our **SMOS** method. The feature extraction backbones f and f_S have an identical structure. f is initialized with ImageNet1K pre-trained weights. **Left:** We first train the Precursor Model $f_S \circ g_S$ to learn scene style shifts with SMD. **Right:** We then perform DG training with training domains from a DG benchmark (e.g. PACS), tuning the DG model $f \circ g$ while being grounded to the SMD-trained Precursor f_S by optimizing $\mathcal{L}_{JS}$.

environments rendered in video games[7, 49, 54], we compile a multi-domain dataset **SuperMarioDomains (SMD)** from video game scenes that can be leveraged as precursor data for Domain Generalization experiments. As its name suggests, our SMD dataset feature synthetic in-game scenes captured from multiple games in the Mario franchise, with image style domains encompass pixelated mosaic 2D graphics in only 50 colors, all the way to high-polygon 3D graphics in 32-bit colors. We land on 4 classes of distinct in-game scenes that consistently appear in all Mario games of our choice. Each scene class has its own unique traits defined by the combination of terrains, objects, and texture. To ensure high consistency in image styles and class labels, we sample our images from video frames of actual game play footage, - neighboring frames of the same game segment share the same scene class. An overview of the samples in SMD is available in Figure 4.

Quantitatively, SMD is designed to counterbalance the stylistic biases of ImageNet1K by having domains that are **low in ICV** and relatively **high in IDD**. On the left of Figure 2, we find that the 4 synthetic-styled domains in SMD have much higher class consistency than ImageNet1K, as in generally lower ICVs. Also, in Figure 3, SMD's domains also keep a series of evenly placed domain shifts in terms of IDD from the dominating style of ImageNet1K, with NES being the most distant domain while N64 and Wii not as close as Photo of PACS or Caltech101 of VLCS.

The SMOS DG Method. Seeing the unique stylistic domain shifts in SMD, we are motivated to design a new DG method with better domain-invariant feature extraction, learning the domain shifts first from isolated class-consistent samples in SMD. We present the **S**cene-grounded **M**inimal d**O**main **S**hift (**SMOS**) method to best leverage the unique domain shifts in SMD. Figure 5 demonstrates the SMOS pipeline. We first train a Precursor Model $\mathcal{M}_S = f_S \circ g_S$ with a Precursor Feature Extractor f_S to learn about the domain shifts of SMD. Sequentially, we train an identically structured model $\mathcal{M} = f \circ g$ with training domains from a DG benchmark, while simultaneously grounded by minimizing the Jensen-Shannon Divergence between distributions of the corresponding feature extractors f_S and f.

Formally, the distribution of the Precursor Model $\mathcal{M}_S$ is learned with the synthetic scene data $\mathcal{D}_S$, namely from SMD. Given N_S stylistic domains of scene image-label pairs $x_j, y_j \in \mathcal{D}_S$, we optimize the cross-entropy loss for updating $\mathcal{M}_S$ in a similar form with ERM:

$$\mathcal{L}_S = \frac{1}{N_S} \sum_{x_j, y_j \in \mathcal{D}_S} \mathcal{L}_{CE}(f_S \circ g_S(x_j), y_j). \quad (7)$$

Together with N training domain images $x_i \in \mathcal{D}_{tr}$, SMOS additionally grounds the training of the DG model $\mathcal{M}$ by minimizing the JSD from the corresponding Precursor Feature Extractor f_S:

$$\mathcal{L}_{JS} = D_{JS}(f_S(x_i), f(x_i)). \quad (8)$$

SMOS optimizes the empirical loss during benchmarking:

$$\mathcal{L}_{SMOS} = \mathcal{L}_S + \mathcal{L}_{ERM} + \lambda \mathcal{L}_{JS}, \quad (9)$$

where the coefficient λ is a hyper-parameter that controls the grounding factor.

6. Experiments and Results

Experiment setups. We use ResNet50 [25] pre-trained with ImageNet1K [13] as the single default backbone for feature extractor networks. Our implementation is based on the source code of DomainBed [23] and MIRO [9]. We evaluate the performance of SMOS on the following DG benchmarks: PACS [36], OfficeHome [57], VLCS [29], TerraIncognita [5], and DomainNet [46].

Two variants of SMOS are designed to initialize $\mathcal{M}_S$ differently - **SMOS**$^-$ whose f_S is initialized from scratch via Kaiming Initialization [24], and **SMOS**$^+$ whose f_S is initialized with ImageNet1K pre-trained weights.

We apply the same data augmentations by ERM [23] and MIRO [9] - each training set image is cropped with random size and aspect ratio, resized to 224×224 pixels, applied random horizontal flips, applied random color jitter, applied grayscale with 10% probability, and normalized using the ImageNet channel means and standard deviations. For hyperparameters, we use a batch size of 16, and the Adam optimizer[30] for all experiments. The hyperparameters listed in Table 4 are consistent with those used in the MIRO paper [9], where for each benchmark, we use a different combination of hyperparameters. We divide all domains in SMD by a 4-to-1 training-test ratio to obtain the

Method	PACS				OfficeHome			DomainNet			
	ACS → Photo	PAS → Cartoon	PAC → Sketch	Avg.	ACP → RealWorld	RAP → Clipart	Avg.	CSIPQ → Real	RSIPQ → Clipart	CRSIP → Quickdraw	Avg.
ERM [56]	97.0±0.1	79.8±1.0	74.7±0.7	84.2	77.4±0.2	53.2±1.2	67.6	60.9±0.1	50.8±0.2	10.5±0.1	44.0
MIRO [9]	97.3±0.2	80.5±0.5	75.9±1.0	85.4	**81.1**±0.4	53.8±0.8	70.5	63.4±0.2	54.2±0.8	11.3±0.0	44.3
SMOS$^-$(ours)	**98.1**±0.0	**84.8**±0.2	81.3±0.4	88.7	80.7±0.3	57.5±0.3	71.0	63.9±0.3	60.3±0.1	13.9±0.0	44.5
SMOS$^+$(ours)	**98.1**±0.1	**84.8**±0.3	**83.2**±0.6	**89.4**	80.8±0.4	**58.6**±0.4	**71.6**	**64.0**±0.2	**62.3**±0.0	**14.7**±0.0	**45.3**

Table 2. DG performance on individual target domains. SMOS gains the greatest improvements on abstract-styled domains while maintaining its performance on photo-styled domains. The letters that precede → denote initial letters of the training domains in the respective benchmarks. The **best results** per setting are shown in bold. Average of 3 trials.

Method	PACS	VLCS	OfficeHome	TerraIncognita	DomainNet	Avg.
Mixstyle† [67]	85.2 ±0.3	77.9 ±0.5	60.4 ±0.3	44.0 ±0.7	34.0 ±0.1	60.3
GroupDRO† [52]	84.4 ±0.8	76.7 ±0.6	66.0 ±0.7	43.2 ±1.1	33.3 ±0.2	60.7
IRM† [33]	83.5 ±0.8	78.5 ±0.5	64.3 ±2.2	47.6 ±0.8	33.9 ±2.8	61.6
ARM† [62]	85.1 ±0.4	77.6 ±0.3	64.8 ±0.3	45.5 ±0.3	35.5 ±0.2	61.7
VREx† [32]	84.9 ±0.6	78.3 ±0.2	66.4 ±0.6	46.4 ±0.6	33.6 ±2.9	61.9
CDANN† [37]	82.6 ±0.9	77.5 ±0.1	65.8 ±1.3	45.8 ±1.6	38.3 ±0.3	62.0
DANN† [17]	83.6 ±0.4	78.6 ±0.4	65.9 ±0.6	46.7 ±0.5	38.3 ±0.1	62.6
RSC† [27]	85.2 ±0.9	77.1 ±0.5	65.5 ±0.9	46.6 ±1.0	38.9 ±0.5	62.7
MTL† [6]	84.6 ±0.5	77.2 ±0.4	66.4 ±0.5	45.6 ±1.2	40.6 ±0.1	62.9
Mixup† [59]	84.6 ±0.6	77.4 ±0.6	68.1 ±0.3	47.9 ±0.8	39.2 ±0.1	63.4
MLDG† [35]	84.9 ±1.0	77.2 ±0.4	66.8 ±0.6	47.7 ±0.9	41.2 ±0.1	63.6
ERM† [56]	84.2 ±0.1	77.3 ±0.1	67.6 ±0.2	47.8 ±0.6	44.0 ±0.1	64.2
SagNet† [43]	86.3 ±0.2	77.8 ±0.5	68.1 ±0.1	48.6 ±1.0	40.3 ±0.1	64.2
CORAL† [53]	86.2 ±0.3	78.8 ±0.6	68.7 ±0.3	47.6 ±1.0	41.5 ±0.1	64.5
CCFP [28]	86.6 ±0.2	78.9 ±0.3	68.9 ±0.1	48.6 ±0.4	41.7 ±0.0	64.8
MIRO† [9]	85.4 ±0.4	79.0 ±0.0	70.5 ±0.4	50.4 ±1.1	44.3 ±0.2	65.9
SMOS$^-$(ours)	88.7 ±0.2	79.7 ±0.1	71.0 ±0.0	**55.5** ±0.8	44.5 ±0.0	67.9
SMOS$^+$(ours)	**89.4** ±0.3	**79.8** ±0.1	**71.6** ±0.1	55.4 ±0.4	**45.3** ±0.0	**68.3**

Table 3. Average leave-one-out cross-validation performances on multiple DG benchmarks. †are baseline results reported in [9]. The **best results** per DG benchmark are highlighted in bold. All accuracies and errors are averaged from 3 trials.

Hyperparameter	PACS	OH	DN	TI	VLCS
Learning rate	3e-5	3e-5	3e-5	3e-5	1e-5
Dropout	0.0	0.1	0.1	0.0	0.5
Weight decay	0.0	1e-6	1e-4	0.0	0.0
Steps	5000	5000	15000	5000	5000
λ	0.15	0.1	0.1	0.1	0.01

Table 4. Hyperparameters for DG experiments. OH, DN, and TI respectively stand for OfficeHome, DomainNet, and TerraIncognita. λ is our grounding coefficient for SMOS as in Equation 9.

precursor model $\mathcal{M}_S$ in SMOS. All experiments are conducted using 2 NVIDIA V100 GPUs.

Domain Generalization Performance with SMOS. We present the most significant improvements of SMOS on the individual target domains in Table 2. The key takeaway is that our SMOS method demonstrates its best strength when generalizing onto more abstract-styled domains, e.g. Sketch in PACS, on which baseline methods such as ERM and MIRO struggle to improve. Using the unique domain shifts in SMD, our SMOS$^+$ variant achieves large improvements over the state-of-the-art MIRO method by $+3.7\%$ and $+7.3\%$ on Cartoon and Sketch in PACS, $+3.6\%$ on Clipart in OfficeHome, and $+8.1\%$ and $+3.4\%$ on Clipart and Quickdraw in DomainNet. We also show that SMOS is able to maintain its performance on photo-realistic domains on par within the MIRO baseline's error range. In some cases, such as PACS and DomainNet, SMOS can even gain slightly better performance than the MIRO baseline.

A comprehensive comparison of a larger selection of DG methods on multiple benchmarks is shown in Table 3. We show that our 2 SMOS variants achieve better overall performance on the chosen DG benchmarks, thanks to the particular improvements on abstract-styled domains. Our SMOS$^+$ variant surpasses the baseline MIRO by $+4.0\%$ on PACS, $+0.9\%$ on VLCS, $+1.1\%$ on OfficeHome, $+4.9\%$ on TerraIncognita, and $+1.0\%$ on DomainNet.

Qualitative Analysis of Domain-invariant Feature Ex-

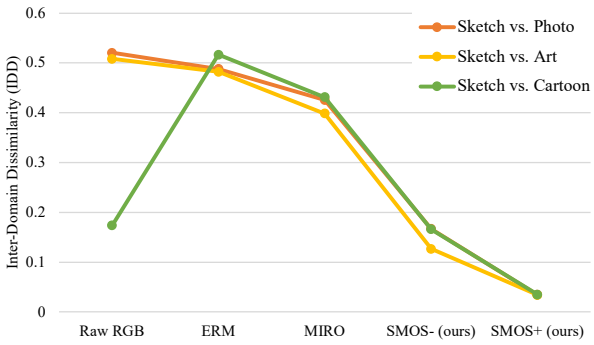


Figure 6. Test vs. training domain IDDs resulted from different DG methods when targeting the Sketch domain of PACS.

Method	Precursor Train. Data $\mathcal{D}_S$	Benchmark Avg. PACS	OfficeHome
ERM	-	84.2	67.6
MIRO	-	85.4	70.5
SMOS$^-$	DomainNet	87.7	65.2
	Places365	87.4	66.5
	SMD (ours)	88.7	71.0
SMOS$^+$	DomainNet	87.9	70.1
	Places365	87.5	66.5
	SMD (ours)	**89.4**	**71.6**

Table 5. DG performance of using substitute precursor data in SMOS$^-$ and SMOS$^+$. All substitute precursor data are downsampled to the same size of SMD (80k). Average of 3 trials.

traction. We further validate SMOS' generalizability, showing it helps project originally distant domains to closer distributional proximity, which qualitatively implicates improvements in extracting domain-invariant features and thus better grounded DG performance. We apply our divergence-based IDD measure between domains that belong to the same benchmark, comparing distributions of representations extracted by different DG methods.

Figure 6 exemplifies the case from the perspective of the Sketch domain in PACS. In raw RGB distributions, Sketch is shown to be highly distant from other domains except Cartoon. When conducting DG image classification tests on Sketch, baseline DG methods such as ERM and MIRO are shown to incrementally lower the distributional divergence from Photo and Art, but at the cost of Cartoon which is originally the most similar to Sketch in RGB. Our methods SMOS$^-$ and SMOS$^+$, along with our SMD precursor data, universally lower the divergences with all non-Sketch domains of PACS. Our methods are thus able to extract domain-agnostic representations within much closer distributional proximity in respective feature spaces, presented as further improved DG performance on Sketch.

Ablation Study of Substitute Precursor Data with SMOS. We also experiment with using substitute image classification datasets as precursor data ($\mathcal{D}_S$ as in Figure 5) under the SMOS method, while SMD is not involved. For substitutes, we choose the benchmark dataset DomainNet [46] which, on its own, has more different domains of image styles and more classes compared to SMD, but also far higher variance in both ICV and IDD than SMD (shown in Figure 1). We also choose the scene-based Places365-Standard (Places365) dataset [66] that has 300+ classes of scenes similar to SMD, but only in one photo-realism style identical to ImageNet1K. We randomly downsample the substitute precursor data to the same size of SMD (80k) at each trial. We use the same 4-to-1 training-test split ratio when obtaining alternative precursor models. We perform benchmarking on PACS and OfficeHome.

Table 5 shows the benchmarking performances of using different precursor data in the SMOS paradigm. We find that SMOS retains the best DG performance when it utilizes SMD's unique stylistic domain shifts and scene labels. In contrast, we see that training the precursor model with ImageNet1K-like image styles (Places365) or less consistent stylistic domains (DomainNet) leads to lower improvements, if not worse performance, than the baseline methods.

7. Conclusions

In this work, we introduce a new paradigm for Domain Generalization (DG) on three facets. We define two new measures, ICV and IDD, to quantitatively understand distributional shifts in stylistic domains based on Jensen-Shannon Divergence. We then present a novel precursor dataset SuperMarioDomains (SMD) sampled from scenes in video games, featuring more consistent categorical classes and dissimilar domains compared to ImageNet1K. We also demonstrate our new DG method SMOS that leverages the unique features of SMD as means to ground the training on DG benchmarks. We find that SMOS along with SMD reaches top performance on multiple DG benchmarks through significant improvements on abstract-styled target domains. SMOS has also been shown to qualitatively improve domain-invariant feature extraction by bringing distant domains within closer divergence in learned feature space. In the future, we would like to explore the application of our methodology on other tasks, such as Controllable Text-to-Image Generation or Visual Question Answering.

Acknowledgements

The authors acknowledge Research Computing at Arizona State University for providing HPC resources and support for this work. This work was supported by NSF RI grants #1750082 and #2132724. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Raphaël Achddou, Yann Gousseau, and Saïd Ladjal. Synthetic images as a regularity prior for image restoration neural networks. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 333–345. Springer, 2021. 2

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 3

[3] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 1

[4] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 2, 3

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 2, 3, 6

[6] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021. 7

[7] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. 2, 6

[8] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. volume 34, page 22405–22418. Curran Associates, Inc., 2021. 3

[9] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 440–457, 2022. 2, 3, 6, 7

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[11] Adrien Courtois, Jean-Michel Morel, and Pablo Arias. Investigating neural architectures by synthetic dataset design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4890–4899, 2022. 2

[12] Steve Dias Da Cruz, Oliver Wasenmuller, Hans-Peter Beise, Thomas Stifter, and Didier Stricker. Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 973–982, 2020. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 6

[14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014. 3

[15] Lixin Duan, Ivor W Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012. 3

[16] Joshua Feinglass and Yezhou Yang. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260, Online, Aug. 2021. Association for Computational Linguistics. 3

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2, 3, 7

[18] 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. *arXiv preprint arXiv:1811.12231*, 2018. 1

[19] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 3

[20] Tejas Gokhale, Joshua Feinglass, and Yezhou Yang. Covariate shift detection via domain interpolation sensitivity. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*, 2022. 3

[21] Goldberger, Gordon, and Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings Ninth IEEE International conference on computer vision*, pages 487–493. IEEE, 2003. 3

[22] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 1, 3

[23] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 3, 6

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[26] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell,

and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013. 3

[27] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020. 7

[28] Xueying Jiang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Domain generalization via balancing training difficulty and model capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18993–19003, 2023. 3, 7

[29] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 2, 3, 6

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2, 3

[32] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 7

[33] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Colorado Reed, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7399–7407, Jun. 2022. 3, 7

[34] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[35] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 7

[36] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2, 3, 6

[37] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 3, 7

[38] Ziyue Li, Kan Ren, Xinyang Jiang, Yifei Shen, Haipeng Zhang, and Dongsheng Li. Simple: Specialized model-sample matching for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3

[39] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. (arXiv:2212.00259), May 2023. arXiv:2212.00259 [cs]. 3

[40] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. 2, 4

[41] Nintendo Co. Ltd. Nintendo game content guidelines for online video & image sharing platforms, Oct 2023. 2

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

[43] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 7

[44] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 1

[45] Goran Paulin and Marina Ivasic-Kos. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial Intelligence Review*, 56(9):9221–9265, Sept. 2023. 2

[46] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2, 3, 6, 8

[47] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018. 3

[48] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12553–12562. IEEE, 2020. 3

[49] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. (arXiv:1608.02192), Aug. 2016. arXiv:1608.02192 [cs]. 3, 6

[50] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 3

[51] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 3

[52] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 7

[53] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frus-

tratingly easy domain adaptation. In *AAAI*, 2016. 2, 7

[54] Geoffrey R. Taylor, Andrew J. Chosak, and Paul C. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. page 1–8, June 2007. 2, 6

[55] Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan A. Plummer. Erm++: An improved baseline for domain generalization. (arXiv:2304.01973), Aug. 2023. arXiv:2304.01973 [cs]. 2, 3

[56] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 2, 3, 7

[57] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2, 3, 6

[58] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5339–5349, 2018. 3

[59] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020. 3, 7

[60] Handi Yu, Simiao Ren, Leslie M. Collins, and Jordan M. Malof. Meta-simulation for the automated design of synthetic overhead imagery. (arXiv:2209.08685), Oct. 2022. arXiv:2209.08685 [cs]. 2

[61] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[62] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23664–23678. Curran Associates, Inc., 2021. 7

[63] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16036–16047, 2023. 3

[64] Lin Zhao and Wenbing Tao. Jsnet: Joint instance and semantic segmentation of 3d point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12951–12958, 2020. 3

[65] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 8

[67] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 7

[68] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. (arXiv:1911.02685), June 2020. arXiv:1911.02685 [cs, stat]. 3