RobustCLEVR: A Benchmark and Framework for Evaluating Robustness in Object-centric Learning

Nathan Drenkow[1,2] Mathias Unberath[1]
[1]The Johns Hopkins University
[2]The Johns Hopkins University Applied Physics Laboratory

Abstract

Object-centric representation learning offers the potential to overcome limitations of image-level representations by explicitly parsing image scenes into their constituent components. While image-level representations typically lack robustness to natural image corruptions, the robustness of object-centric methods remains largely untested. To address this gap, we present the RobustCLEVR benchmark dataset and evaluation framework. Our framework takes a novel approach to evaluating robustness by enabling the specification of causal dependencies in the image generation process grounded in expert knowledge and capable of producing a wide range of image corruptions unattainable in existing robustness evaluations. Using our framework, we define several causal models of the image corruption process which explicitly encode assumptions about the causal relationships and distributions of each corruption type. We generate dataset variants for each causal model on which we evaluate state-of-the-art object-centric methods. Overall, we find that object-centric methods are not inherently robust to image corruptions. Our causal evaluation approach exposes model sensitivities not observed using conventional evaluation processes, yielding greater insight into robustness differences across algorithms. Lastly, while conventional robustness evaluations view corruptions as out-of-distribution, we use our causal framework to show that even training on in-distribution image corruptions does not guarantee increased model robustness. This work provides a step towards more concrete and substantiated understanding of model performance and deterioration under complex corruption processes of the real-world.[1]

1. Introduction

Common deep neural network (DNN) architectures have been shown to lack robustness to naturally-induced image-level degradation [12, 20, 46, 48]. In safety-critical scenarios, any reduction in model performance due to naturally-occurring corruptions poses a threat to system deployment. While many proposed solutions exist for increasing the robustness of image-level representations [4,7,9,21,33,41,49, 51], a measurable performance gap remains [14, 48].

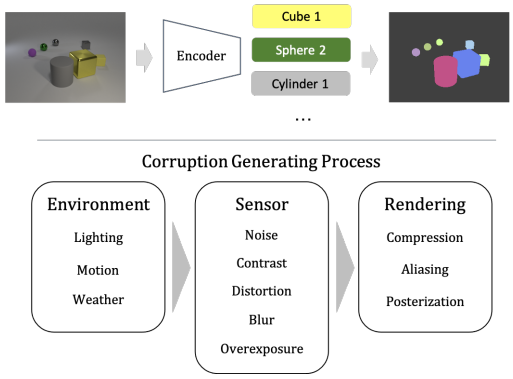


Figure 1. (Top) Object-centric methods explicitly parse scenes into constituent objects. (Bottom) The corruption generating process involves with many causal factors with complex dependencies.

Recent advances in object-centric (OC) representation learning signal a paradigm shift towards methods that explicitly parse visual scenes as a precursor to downstream tasks. These methods offer the potential to analyze complex scene geometries, support causal reasoning, and reduce the reliance of deep learning models on spurious image features and textures. Recent works have examined the use of OC representations for downstream tasks [11] and action recognition [50] showing positive benefits of such representations over more traditional image-level features. One hypothesis is that OC methods inherently learn scene-parsing mechanisms which are tied to stable features of the scene/objects and robust to naturally-induced image corruptions. However, quantitative proof of these desirable properties has not yet been obtained. We test this hypothesis by conducting the first analysis of the robustness of OC representation learning to non-adversarial, naturally-induced image corruptions.

[1]Data and code will be released upon publication

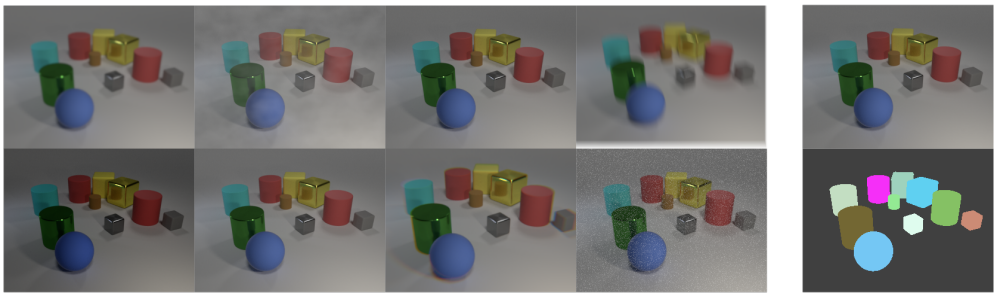


Figure 2. **RobustCLEVR image corruptions** rendered independently from left-to-right: (Top) Blur, cloud, defocus, displacement blur, (Bottom) Gamma, glare, lens distortion, noise. Rightmost column is the clean image and ground truth mask.

1.1. Background

Robustness evaluation: Robustness to common corruptions has been previously addressed in a number of other contexts [20, 25, 31, 38, 39]. However, prior work has made several strong limiting assumptions which we aim to address here. First, prior work has treated categories of image corruptions as IID, failing to account for causal relationships in the image generation process (e.g., low brightness causes longer exposure times or higher sensor sensitivity, resulting in motion artifacts or increased quantum noise, respectively). The lack of interactions leads to a set of image corruptions potentially decoupled from reality.

Second, corruption severity is often modeled heuristically without controlling for the impact on image quality and assuming all severities are equally likely. Defining corruption severity on a discrete scale [20] often ignores the fact that severity is continuous in real-world conditions (e.g., blur due to motion depends on the velocity of the system or scene). Furthermore, because DNNs are highly non-linear, performance change due to severity is also likely non-linear. Since corruption severity in the real world is often non-uniform, robustness evaluations should reconsider the nature of the assumed severity distribution.

Lastly, prior work typically assumes common corruptions are out-of-distribution (OOD), positing that they are not actually "common" (or even present) within the training sample distribution. While this has benefits for assessing model performance on unseen conditions, it fails to consider the more likely scenario where the assumed distribution naturally contains image corruptions (even if rare) and the model may have access to corrupted samples during training. Even in that case, the evaluation can still target specific corruption conditions while ensuring that the model has also been trained on data representative of the assumed "true" distribution.

OC robustness evaluation: OC representation learning is formulated as an unsupervised object discovery problem, so the absence of annotations (e.g., semantic labels, bounding boxes, object masks) forces models to learn only from the structure and imaging conditions inherent to the data distribution. To successfully develop OC methods that work on highly variable real-world data, robustness evaluations must be also able to account for a wider set of imaging conditions consistent with the image generation process. Training and evaluating the robustness of OC methods thus relies on a clear statement of the assumptions underlying the train and test distributions.

Our approach and contributions: We address limitations of prior robustness work by developing a causal inference framework that enables us to unify common interpretations of robustness while also providing a means to improve the alignment between evaluation data and challenging, domain-specific imaging conditions. We show that knowledge of the image generation process enables the specification of causal graphs which explicitly capture assumptions about the sources of and dependencies between image corruptions (described further in Section 3.3). We conduct the first extensive evaluation of the robustness of OC methods using contrasting causal models of the data generating process (DGP) and show that assumptions about the causal model structure and its underlying distribution are critical for interpreting OC model robustness. Lastly, we demonstrate that common corruption robustness of OC methods can also be interpreted as dealing with long-tail image distributions contrary to the more restrictive OOD assumption.

To investigate the robustness of OC methods, we developed the **RobustCLEVR** framework and dataset. We build off prior works which initially evaluated OC methods on CLEVR [24] and CLEVRTex [26], datasets consisting of a collection of uncorrupted scenes composed of sets of simple objects with varying complexity of color, material, and texture properties. We use our causal framework to generate multiple variants of RobustCLEVR with each investigating the effects of distributional assumptions on measured model robustness. This benchmark and framework is the first of its kind for OC learning and provides a stepping stone towards realizing its potential on real-world data.

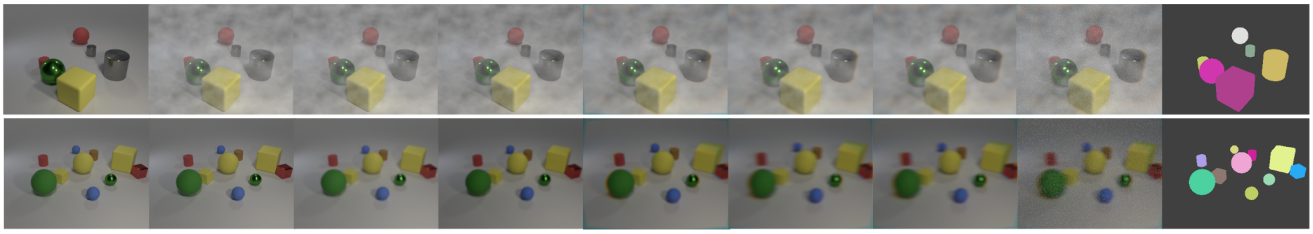


Figure 3. RobustCLEVR variant with causally-dependent corruptions. Rows are different samples from the same causal model and columns are images rendered at each node of the model. Corruptions are rendered according to order of the causal model from left to right: Clean, cloud, blur, gamma, lens distortion, displacement/motion blur, defocus blur, noise, ground truth.

2. Related Work

Robustness benchmarks Robustness in deep learning for computer vision has been extensively studied outside of OC learning [14]. Several benchmarks have enabled systematic evaluation of robustness of deep learning methods with respect to image classification [20, 31, 32, 38], object detection [13, 39], instance segmentation [1], and distribution shifts [29, 54]. While challenging datasets for OC learning such as CLEVRTex [27] have helped push the boundaries of these methods, datasets for evaluating robustness to image corruption remains unexplored.

Object-centric methods Methods for decomposing scenes into their constituent objects (without explicit object-level labels) have emerged recently under a variety of names including unsupervised object discovery, unsupervised semantic segmentation, and OC learning. Early techniques [8, 18, 30] processed images via a series of glimpses and learned generative models for scene construction by integrating representations extracted over multiple views. More recent techniques learn models capable of generating full scenes from representations bound to individual objects. For generative [3, 16, 17, 19, 23, 34] and discriminative [35] methods, image reconstruction plays a crucial role in the learning objective.

Beyond static scene images, multi-view and video datasets provide additional learning signals for unsupervised object discovery. Recent techniques [2, 15, 26, 28, 47] have exploited object motion estimated via optical flow for improving OC representations. In contrast, multi-view techniques [42, 43] take advantage of overlapping camera perspectives for improving scene decomposition. We focus on static scenes in this work and multi-view methods remain candidates for future evaluation.

Causal inference for robustness Lastly, causal inference and computer vision research have become increasingly intertwined in recent years [5, 14, 45]. Early works [6, 36] focused on causal feature learning and have expanded to other vision tasks and domains [22, 37, 44, 53]. Causal inference and robustness have also been explored in the context of adversarial [53] and non-adversarial [10, 40, 52] conditions.

3. Methods

3.1. Structural causal models

Structural causal models (SCM) consist of variables, their causal relationships, and their distributional assumptions, all of which describe an associated data generating process. The DGP can be represented as a Directed Acyclic Graph (DAG) $\mathcal{G}$ consisting of variables ($\mathcal{V}$) as nodes and relationships ($\mathcal{E}$) as edges and where the output of the node is a function of its parents and an exogenous noise term. A joint distribution over all variables underlies the SCM which encodes their dependencies.

In computer vision, knowledge of the imaging domain and vision task provides a means for constructing such SCMs. While full knowledge of $\mathcal{V}$, $\mathcal{E}$, and distributional information is generally not possible, plausible SCMs of the data generating process may still be constructed. These SCMs encode expert knowledge and assumptions which can be verified through observational data. Alternatively, in simulated data, as in the case of CLEVR, we have full knowledge of the DGP including access to all variables and their underlying distributions. Critically, our framework allows us to leverage this access to fully specify *arbitrary* graphical causal models of the DGP and then generate data in accordance with those models and their underlying distributions.

3.2. Robustness

To date, the definition of robustness in computer vision has assumed many forms [14] including (but not limited to) adversarial or worst-case behavior, out-of-distribution performance, and domain generalization. We aim to unify many of these interpretations via a causal framework.

First, we make a key distinction: the SCM of the data generating process describes our *a priori* beliefs about the true data distribution, independent of any sampling of the data. This provides a frame of reference for specifying robustness conditions such as when image corruptions are rare, due to distribution shift, or out-of-distribution. When evaluating robustness, we rely on this distinction in order to verify that the sampled training and evaluation datasets are

consistent with our assumptions about the true underlying DGP.

Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the structural causal model of a data generating process. In the case of natural images, the nodes $\mathcal{V} = \{v_i\}$ represent variables such as the concepts of interest, distractor concepts, environmental conditions, and sensor properties. The model $\mathcal{G}$ induces a joint distribution $p_G(\{v_i\})$ over all variables $\mathcal{V}$ where $p(v_1, \ldots, v_i) = \Pi_{j \leq i} p(v_j | pa_j)$ where pa_j are the parents of j in $\mathcal{G}$.

We consider common perspectives of robustness conditions in the context of structural causal models as follows.

- **Distribution shift** - Any shift in the marginal or conditional distributions of nodes in $\mathcal{V}$.
- **Out-of-distribution (OOD)** - The case when test concepts or image conditions are not in the support of $p_{\mathcal{G}}(\{v_i\})$. This can be viewed as a special and extreme case of distribution shift.
- **Long-tail robustness** - Samples drawn from the DGP which are rare relative to the joint, marginal, and/or conditional distributions.
- **Adversarial** - Direct image manipulation performed via intervention (i.e., $do(X = x + \delta)$ where $do(\cdot)$ represents setting the value of X independent of its parents and δ represents the adversarial perturbation added to the clean image)

This framework naturally allows for precise definition of the known/assumed robustness conditions as they relate to specific nodes of the DGP, which is in contrast to many common approaches which paint robustness in broad strokes. Conventional robustness evaluations are still included as a special case (i.e., IID corruptions assumed to be OOD) while more general evaluations may be implemented via soft/hard interventions on any subset of nodes in the SCM/DAG. These interventions measure the effects of specific types of corruptions on the image generating process by manipulating node values/distributions (independent of their parents) while maintaining downstream causal relationships.

3.3. SCM of the Corruption Generating Process

For studying the robustness of OC methods, we consider the case where image scenes composed of a finite set of objects are corrupted according to various imaging conditions. Objects and scene geometry are first sampled independent of imaging conditions so that we can focus our attention on modeling the corruption generation process. We define an SCM/DAG $\mathcal{G} = (\mathcal{C}, \mathcal{E})$ where each c_i applies a corruption to the already-constructed scene and edges e_{ij} indicate dependencies between corruption types.

For each corruption, we associate one (or more) severity parameter γ_i such that for any image x, corruption c, and similarity metric $m(x, c(x; \gamma))$, we observe greater image degradation as γ increases:

$$\gamma_i > \gamma_j \Rightarrow m(x, c(x; \gamma_i)) < m(x, c(x; \gamma_j))$$
$$c(x; \gamma = 0) = x$$

Severity parameters are sampled from the causal model such that $\gamma_i = f_i(\gamma_{pa_i}, \epsilon_i)$ where the causal mechanism f_i is a function of γ_{pa_i}, the severity parameters for the parents of node i, and ϵ_i, a noise term.

3.4. Generating RobustCLEVR

The RobustCLEVR framework supports the definition of arbitrary SCMs/DAGs which capture various structural relationships and distributional assumptions regarding the data generating process and corresponding image corruptions. Image corruptions are implemented via Blender workflows or "recipes" and are typically defined by one or a few corruption parameters (i.e., the γ_i from Sec. 3.3). While arbitrary corruption recipes may be defined using Blender to achieve a range of photorealistic effects, we implemented Gaussian blur, defocus blur, displacement/motion blur, gamma, clouds, white noise, glare, and lens distortion.

For each image, the initial set of objects, their materials, and their placement in the scene are first sampled according to [24]. We then sample corruption parameters from the distribution defined by the SCM/DAG. Using the Blender Python API, we apply the corruptions (given their sampled parameters) to the scene according to the ordering specified by the DAG.

Our framework evaluates robustness in two novel ways. (1) Specification of the SCM/DAG allows for the generation of a wide range of *unseen distortions* that may result from complex interdependencies/relationships between corruptions. (2) Unlike prior works which consider corruption severity only at discrete and heuristic levels, samples from a DGP defined in our framework have corruption severities which vary continuously and consistent with the underlying distribution. Crucially, (1) and (2) enable better alignment with real world conditions where types of image corruptions rarely occur in pure isolation and their impact on image quality is continuously varying.

4. Experiments and Results

Baseline Algorithms We evaluate pixel- and glimpse-based OC algorithms for all experiments. For pixel-based methods we evaluate EfficientMORL [16], GENESISv2 [17], and IODINE [19]. For glimpse-based methods, GNM [23], SPACE [34], and SPAIR [8] are evaluated. With the exception of Experiment 4 (Sec. 4.4), all models were trained on the public CLEVR training set and evaluated on the appropriate RobustCLEVR variants. We use code for baseline algorithms originally provided by [27].

Table 1. Mean Intersection over Union by model for IID-sampled corruptions. Rows within groups correspond to whether the corruption severity is sampled uniformly. Highlighted cells indicate the best performance in that column.

Model	Severity	mIoU Blur		Clouds		Defocus		Gamma		Lens Distortion		Motion Blur		Noise		Clean	
GENESISv2	Non-uniform	38.67	±0.31	35.70	±0.36	39.04	±0.31	28.86	±0.39	18.64	±0.25	22.13	±0.35	39.25	±0.31	38.94	±0.31
	Uniform	39.24	±0.31	38.35	±0.32	38.93	±0.31	26.01	±0.40	26.27	±0.28	30.27	±0.33	39.54	±0.31	39.00	±0.31
GNM	Non-uniform	52.77	±0.58	29.98	±0.77	58.41	±0.52	51.75	±0.75	27.13	±0.52	24.88	±0.65	56.71	±0.53	61.32	±0.50
	Uniform	56.38	±0.52	35.37	±0.72	54.47	±0.54	45.50	±0.76	43.14	±0.50	40.08	±0.63	58.01	±0.51	61.01	±0.50
IODINE	Non-uniform	63.75	±0.45	32.83	±0.96	66.60	±0.42	27.84	±0.77	26.83	±0.46	30.78	±0.64	65.84	±0.45	66.20	±0.40
	Uniform	65.77	±0.42	39.51	±0.94	64.16	±0.43	22.78	±0.70	42.11	±0.47	46.14	±0.61	67.13	±0.42	66.24	±0.40
SPACE	Non-uniform	42.98	±0.69	31.85	±0.70	49.54	±0.63	45.26	±0.70	17.28	±0.40	19.51	±0.53	49.23	±0.63	51.09	±0.64
	Uniform	48.37	±0.62	37.65	±0.67	45.37	±0.63	42.95	±0.72	28.36	±0.48	30.10	±0.62	49.85	±0.63	50.95	±0.63
SPAIR	Non-uniform	69.74	±0.56	49.04	±0.86	71.86	±0.56	30.85	±0.91	23.60	±0.49	31.19	±0.68	71.13	±0.57	72.99	±0.57
	Uniform	71.22	±0.55	57.88	±0.67	69.92	±0.54	25.16	±0.82	39.60	±0.58	46.99	±0.68	72.13	±0.56	73.04	±0.56
eMORL	Non-uniform	18.05	±0.24	14.23	±0.24	18.47	±0.25	14.62	±0.22	12.05	±0.17	12.47	±0.19	17.86	±0.24	18.54	±0.25
	Uniform	20.42	±0.34	20.60	±0.29	20.04	±0.34	15.46	±0.23	15.31	±0.24	16.63	±0.29	20.01	±0.32	20.67	±0.34

Table 2. Mean Squared Error (MSE) by model for corruptions sampled IID. Lower MSE indicates better recovery of the original clean image. Highlighted cells indicate the best performance in that column.

Model	Severity	MSE Blur		Clouds		Defocus		Gamma		Lens Distortion		Motion Blur		Noise		Clean	
GENESISv2	Non-uniform	58.60	±2.06	193.36	±6.33	38.19	±1.03	216.07	±7.13	599.75	±11.79	402.55	±9.64	38.10	±1.06	26.62	±0.68
	Uniform	40.22	±1.13	131.23	±3.88	50.49	±1.55	259.92	±6.41	324.83	±7.27	230.88	±6.14	31.35	±0.79	26.80	±0.68
GNM	Non-uniform	117.60	±2.63	288.40	±7.31	96.39	±1.93	875.12	±36.45	598.87	±12.20	390.67	±8.81	103.53	±2.08	87.10	±1.79
	Uniform	100.12	±2.03	230.11	±5.60	107.94	±2.23	1182.75	±36.29	319.91	±7.05	246.39	±5.55	95.93	±1.92	87.42	±1.78
IODINE	Non-uniform	76.52	±2.16	455.67	±13.47	58.05	±1.51	1720.36	±55.11	592.79	±12.05	382.55	±8.47	88.62	±2.58	49.76	±1.36
	Uniform	60.74	±1.55	381.64	±12.13	69.67	±1.81	2246.86	±56.38	318.31	±7.05	229.03	±5.62	66.78	±1.59	50.05	±1.37
SPACE	Non-uniform	98.27	±2.36	325.26	±10.11	76.14	±1.59	788.30	±19.61	591.01	±12.12	367.83	±8.29	89.43	±1.88	62.83	±1.33
	Uniform	80.95	±1.74	221.29	±5.85	90.13	±2.04	950.58	±19.44	319.09	±7.00	220.79	±5.19	78.04	±1.59	63.13	±1.36
SPAIR	Non-uniform	81.14	±2.16	770.07	±32.68	63.16	±1.55	1797.57	±59.94	623.68	±12.88	408.85	±9.15	90.24	±2.48	53.18	±1.41
	Uniform	66.23	±1.64	381.76	±11.76	74.88	±1.87	2377.58	±61.43	334.74	±7.39	247.47	±6.13	69.51	±1.68	53.60	±1.43
eMORL	Non-uniform	62.58	±2.04	1044.46	±51.85	42.03	±1.09	2014.28	±58.91	609.37	±12.68	371.51	±8.61	85.20	±2.92	32.11	±0.84
	Uniform	46.85	±1.29	455.46	±14.41	58.75	±1.70	2662.18	±63.08	310.45	±6.91	211.35	±5.40	51.65	±1.24	31.78	±0.91

Metrics Consistent with prior work, we report performance on mean Intersection over Union (mIoU) and Mean Squared Error (MSE). The mIoU metric measure the ability of the model to locate and isolate individual objects in the scene (i.e., object recovery) while MSE measures reconstruction quality (i.e., image recovery). Metrics are computed relative to the uncorrupted images and the corresponding masks. For each baseline in Experiments 1-3, we train a set of three models corresponding to different random seeds. Due to significant variability in mIoU, we report metrics for the model+seed with the highest clean mIoU and obtain confidence intervals using 1000 bootstrap samples of predictions for each corruption. These results represent an upper bound on performance.

Dataset variants In each experiment, a test set is generated consisting of 10k distinct scenes. For each scene, corruptions are rendered according to the parameters and ordering determined by the associated causal model. For eight corruption types and 10k scenes, this yields 80k images for evaluation per experiment. Within each scene object properties including type, color, material, scale, and placement are all randomized according to the original CLEVR dataset described in [24].

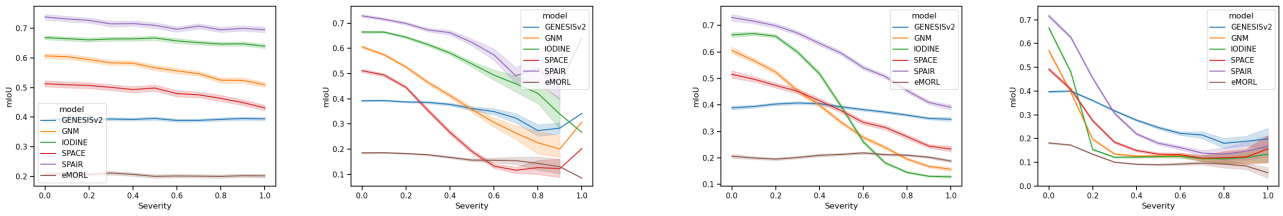
4.1. Experiment 1: Independent corruptions, uniform severity

We first generate a RobustCLEVR variant where the causal model produces IID corruptions (where corruption parameters γ are sampled uniformly - See Appendix for distribution details). This corresponds to the conventional corruption evaluations where corruptions are OOD and independent with severity uniformly distributed. Results of evaluating OC methods on this data are found in Tables 1 and 2

The results indicate that the ability to recover underlying objects is largely tied to the distribution of corruption severity across the different corruption types. Figure 4 shows how mIoU differs as a function of severity for each algorithm. For instance, for the cloud corruption, we see that SPAIR and IODINE report similar mIoU at low severity but SPAIR's performance degrades more gracefuly as severity increases.

4.2. Experiment 2: Independent corruptions, non-uniform severity

We next examine the impact of independent corruptions with non-uniform severity. A similar RobustCLEVR variant is generated with the same DAG as Experiment 1 but where the corruption parameter(s) for each node are sampled from non-uniform distributions. Since for most pa-



(a) Blur corruption (b) Cloud corruption

Figure 4. Object recovery (mIoU) as a function of normalized severity. The severity is calculated by normalizing the sampled corruption parameter distribution to the interval $[0, 1]$ (with each panel normalized independently). For each corruption (panels (a), (b)), severity is sampled (left) uniformly and (right) non-uniformly.

Table 3. Mean Intersection over Union (mIoU) by model for corruptions sampled non-IID. Corruption order in the table from left to right reflects the sampling order in the causal model. Higher mIoU indicates better recovery of the original clean image. Highlighted cells indicate the best performance in that column.

		mIoU															
Model	Severity	Clouds		Blur		Gamma		Lens Distortion		Motion Blur		Defocus		Noise		Clean	
GENESISv2	Non-uniform	38.75	±0.30	38.88	±0.30	36.88	±0.31	34.32	±0.31	31.79	±0.29	31.80	±0.29	32.28	±0.29	38.86	±0.30
	Uniform	38.18	±0.32	38.33	±0.32	34.71	±0.34	31.98	±0.31	29.85	±0.29	29.88	±0.29	30.61	±0.28	38.90	±0.31
GNM	Non-uniform	55.02	±0.73	52.95	±0.71	55.54	±0.70	53.71	±0.67	48.78	±0.65	48.42	±0.65	47.13	±0.66	61.32	±0.50
	Uniform	53.87	±0.77	50.81	±0.74	53.29	±0.75	51.18	±0.71	47.04	±0.69	45.56	±0.67	44.68	±0.69	60.99	±0.51
IODINE	Non-uniform	59.87	±0.74	59.56	±0.75	49.88	±0.85	46.49	±0.78	41.34	±0.66	41.16	±0.67	41.75	±0.69	66.56	±0.41
	Uniform	58.51	±0.81	58.06	±0.81	39.41	±0.91	36.57	±0.81	32.59	±0.66	31.85	±0.66	32.85	±0.69	66.40	±0.40
SPACE	Non-uniform	47.92	±0.67	46.76	±0.66	47.69	±0.66	42.95	±0.62	37.56	±0.61	37.36	±0.60	36.31	±0.60	51.31	±0.63
	Uniform	46.49	±0.72	44.78	±0.70	45.76	±0.69	40.89	±0.64	36.51	±0.63	35.17	±0.61	33.74	±0.60	51.07	±0.62
SPAIR	Non-uniform	69.60	±0.66	68.73	±0.65	58.70	±0.89	53.24	±0.82	46.53	±0.69	46.37	±0.70	47.14	±0.69	73.28	±0.57
	Uniform	67.35	±0.76	66.13	±0.74	46.83	±1.01	41.86	±0.87	36.70	±0.70	36.12	±0.71	36.99	±0.72	72.93	±0.57
eMORL	Non-uniform	17.98	±0.25	17.92	±0.25	17.12	±0.25	16.46	±0.24	15.61	±0.22	15.54	±0.22	15.39	±0.22	18.60	±0.26
	Uniform	17.59	±0.26	17.48	±0.26	15.78	±0.24	15.22	±0.22	14.57	±0.21	14.33	±0.21	14.19	±0.21	18.60	±0.25

Table 4. Mean Squared Error (MSE) by model for corruptions sampled non-IID. Corruption order in the table from left to right reflects the sampling order in the causal model. Lower MSE indicates better recovery of the original clean image. Highlighted cells indicate the best performance in that column.

		MSE															
Model	Severity	Clouds		Blur		Gamma		Lens Distortion		Motion Blur		Defocus		Noise		Clean	
GENESISv2	Non-uniform	52.49	±2.70	57.66	±2.75	80.43	±3.15	124.71	±3.60	171.34	±3.90	172.97	±3.91	166.59	±3.90	26.62	±0.69
	Uniform	64.80	±4.12	73.34	±4.08	120.74	±4.24	168.43	±4.27	207.51	±4.33	212.16	±4.42	200.17	±4.53	26.62	±0.67
GNM	Non-uniform	122.31	±4.03	127.52	±4.04	190.73	±9.72	217.21	±9.77	252.33	±9.53	251.75	±9.41	235.50	±7.98	87.21	±1.82
	Uniform	133.39	±5.09	141.30	±5.04	289.28	±12.18	317.00	±11.81	346.80	±11.37	346.07	±11.25	302.66	±9.02	87.25	±1.77
IODINE	Non-uniform	131.41	±8.26	137.09	±8.27	382.47	±22.97	417.21	±22.81	457.00	±22.06	458.36	±22.04	419.13	±19.75	50.02	±1.39
	Uniform	143.15	±9.34	151.12	±9.35	683.71	±30.12	720.13	±29.58	752.83	±28.77	761.58	±28.64	673.68	±24.93	49.79	±1.33
SPACE	Non-uniform	101.92	±4.10	109.77	±4.11	239.08	±10.77	273.42	±10.78	309.49	±10.21	309.46	±10.14	294.86	±9.41	62.77	±1.35
	Uniform	123.10	±6.43	134.88	±6.36	401.79	±14.03	439.68	±13.84	469.27	±13.06	468.05	±12.73	440.53	±11.80	63.13	±1.35
SPAIR	Non-uniform	134.54	±8.15	140.47	±8.12	371.67	±22.71	409.36	±22.45	454.70	±21.89	455.62	±21.92	419.69	±19.69	53.34	±1.40
	Uniform	216.09	±18.82	224.80	±18.73	721.41	±31.21	759.17	±30.44	797.32	±29.55	803.72	±29.72	726.93	±27.35	53.23	±1.38
eMORL	Non-uniform	147.29	±11.24	153.64	±11.29	469.62	±26.98	503.31	±26.72	545.94	±26.04	544.75	±25.84	502.95	±23.45	32.26	±0.84
	Uniform	258.83	±27.74	268.29	±27.84	926.02	±39.64	959.81	±39.23	993.88	±38.18	991.45	±37.84	904.73	±36.42	32.27	±0.85

rameters, monotonically increasing/decreasing the value of a corruption parameter corresponds to an increase in the severity of the corruption, the uniform distribution from Experiment 4.1 is replaced with a Half-Normal distribution. This trades off a bias towards low-severity cases with the possibility of sampling higher severity cases from the distribution tails. Images from this variant are visualized in Figure 2. Results of evaluating OC methods on this data are found in Tables 1 and 2.

Results show that long-tailed severity distributions lead to measurable changes in absolute and relative values of mIoU across models. For instance, performance generally improves for the Gamma corruption when the severity distribution shifts from uniform to non-uniform, whereas Lens Distortion or Motion Blur exhibit lower performance as a result of the long-tail. Furthermore, Figure 4 also illustrates non-linear relationships between performance and severity.

4.3. Experiment 3: Dependent corruptions

An important benefit of the causal graph relative to the current standard robustness evaluation approach is the ability to describe causal relationships known or assumed to exist in the image domain of interest. As such, we next consider a more challenging RobustCLEVR variant where the

underlying causal model follows a chain structure. Corruptions are linked sequentially and sampled corruption parameters are a function of the parameter values of their immediate parent.

As in Experiment 4.2, we also consider the impact of distributional assumptions on the measured robustness. We create an additional variant of the chain model with non-uniform severity distributions and evaluate the performance of OC methods on the data generated from this model as well. While the causal model variants in this experiment no longer produce IID corruptions as in Experiment 1 and 2, the evaluation is still considered OOD since all models were trained on only clean CLEVR data. Results for Experiment 3 are found in Tables 3 and 4.

While the chain DAG structure suggests that the total image corruption increases as images are sampled in sequence along the DAG, the causal mechanisms and distributions at each node also dictate how each corruption severity is sampled. For instance, this may lead to larger differences in performance from one corruption to the next in the model (e.g., Blur $\rightarrow$ Gamma vs. Defocus $\rightarrow$ Noise).

4.4. Experiment 4: Long-tail Robustness

Lastly, many real world scenarios allow for the possibility that corrupted images are in-distribution (ID) but occur infrequently in the training set (either due to the rarity of the corruption in reality or due to sampling bias such as preferences by annotators for labeling clean images). We generate a RobustCLEVR variant which treats the corruptions as in-distribution but occurring with low probability. As in Section 4.2, the causal model DAG is specified as a tree of depth 1 whereby all corruptions are mutually independent and severities are non-uniformly distributed (see Appendix for details). For each OC method, we train two models, one on only clean data and one on data including corruptions with $p_{corr} = 0.01, p_{clean} = 1 - \sum_i p_{corr_i}$. Each training dataset consists of 50k unique scenes.

All models are evaluated on a separate corrupted test set sampled from the same causal model used for training as well as the test set from Experiment 1 which contained IID corruptions with uniform severity. These test sets correspond to the long-tail robustness and distribution shift cases described in Section 3.2. Results are shown in Table 5.

With all models (excluding SPACE), the inclusion of corrupted samples in the training set appears to generally decrease robustness for the corresponding model. For models like GNM and GENESISv2, the performance differences are small whereas models like IODINE and eMORL often differ by $> 10\%$ when corruptions are included/excluded from the training set. These trends hold for evaluation on both the uniform and non-uniform distributions for severity. This is discussed in more detail in Section 5.

5. Discussion

The experiments in Section 4 suggest that OC methods are not immune to image corruptions. While it is not surprising that performance degradation would occur in these cases, the sensitivity to low-severity corruptions suggests that OC models are not inherently more robust than non-OC techniques. We attribute much of this finding to the use of image reconstruction as a common component of the learning objective for these models. For generative methods, this is due to the log likelihood term in the ELBO objective while discriminative methods like Slot Attention use MSE directly. Consistent with results on CLEVRTex [27], models which produce lower MSE (i.e., better image recovery) also tend to produce lower mIoU (i.e., object recovery). The use of image reconstruction by OC models during learning may encourage the latent representations to encode nuisance or appearance factors not critical to scene parsing. The result of this learning strategy is poor object recovery when those same nuisance factors are modified or corrupted as in a robustness scenario.

We also find that the structure and corresponding distribution of the underlying data generating process matters in assessing model robustness. We observe measurable performance differences as a result of changing causal and distributional assumptions. For instance, considering two top models from Experiments 1-3, GNM and SPAIR, we observe differences in relative mIoU performance on the same set of corruptions drawn from the IID (Table 1) and non-IID causal models (Table 3). While we expect the mIoU to change for each model as a result of the distribution shift, the disparity in mIoU between the two models for any given corruption is not constant between the IID and non-IID scenarios. When causal models are defined to approximate specific real-world distributions, measuring such performance differences may be critical to understanding and predicting model behavior in the wild.

Lastly, the results of Experiment 4 indicate that robustness is not a purely OOD problem. The inclusion of corrupted data as rare samples in the training distribution has a negative impact on robustness for many of the models. This warrants further research as it contradicts existing findings for robustness in supervised, discriminative models where data augmentation with heavy corruption or other image transformations yields significant gains in robustness to common corruptions [7, 9, 21, 33, 41, 49, 51]. One possible explanation is that the corrupted images (while rare in training), simply provide less informative signal about the scene geometry and object properties. Alternatively, when the training sample size is fixed, the inclusion of these corrupted images also means that fewer clean images are also available for learning. When corrupted images are in distribution, OC models with image reconstruction objectives may be increasingly incentivized to reconstruct low level

Table 5. Comparison of model performance when corruptions with non-uniform severity are in-distribution (clean + corrupt) and out of distribution (clean only). Highlighted cells indicate the best performance in that column.

Model	Train Distribution	Severity	mIoU							
			Blur	Clouds	Defocus	Motion Blur	Gamma	Lens Distortion	Noise	Clean
GENESISv2	clean	Non-uniform	0.218	0.201	0.224	0.137	0.204	0.130	0.224	0.225
	clean + corrupt	Non-uniform	0.191	0.185	0.193	0.147	0.196	0.121	0.189	0.194
	clean	Uniform	0.224	0.217	0.221	0.179	0.200	0.167	0.226	0.225
	clean + corrupt	Uniform	0.193	0.188	0.192	0.167	0.196	0.153	0.192	0.195
GNM	clean	Non-uniform	0.466	0.243	0.524	0.226	0.493	0.247	0.489	0.550
	clean + corrupt	Non-uniform	0.456	0.231	0.515	0.221	0.367	0.246	0.487	0.543
	clean	Uniform	0.503	0.279	0.485	0.359	0.423	0.390	0.512	0.548
	clean + corrupt	Uniform	0.494	0.262	0.475	0.352	0.297	0.387	0.504	0.540
IODINE	clean	Non-uniform	0.627	0.303	0.651	0.310	0.275	0.262	0.628	0.647
	clean + corrupt	Non-uniform	0.274	0.210	0.284	0.180	0.290	0.155	0.277	0.288
	clean	Uniform	0.645	0.356	0.635	0.463	0.230	0.412	0.649	0.649
	clean + corrupt	Uniform	0.283	0.233	0.279	0.228	0.275	0.207	0.281	0.288
SPACE	clean	Non-uniform	0.123	0.123	0.123	0.123	0.123	0.123	0.123	0.123
	clean + corrupt	Non-uniform	0.664	0.386	0.717	0.273	0.617	0.255	0.708	0.732
	clean	Uniform	0.123	0.123	0.123	0.123	0.123	0.123	0.123	0.123
	clean + corrupt	Uniform	0.704	0.472	0.683	0.463	0.554	0.434	0.716	0.730
SPAIR	clean	Non-uniform	0.685	0.475	0.706	0.309	0.290	0.232	0.693	0.716
	clean + corrupt	Non-uniform	0.682	0.578	0.700	0.315	0.624	0.230	0.697	0.708
	clean	Uniform	0.701	0.559	0.688	0.463	0.241	0.389	0.705	0.716
	clean + corrupt	Uniform	0.694	0.640	0.682	0.463	0.615	0.388	0.703	0.707
eMORL	clean	Non-uniform	0.397	0.318	0.406	0.226	0.204	0.205	0.421	0.411
	clean + corrupt	Non-uniform	0.192	0.170	0.196	0.130	0.219	0.120	0.184	0.197
	clean	Uniform	0.405	0.384	0.400	0.313	0.181	0.283	0.420	0.410
	clean + corrupt	Uniform	0.197	0.179	0.195	0.161	0.204	0.143	0.185	0.199

corruptions which have no bearing on object recovery. So while OC methods aim to represent objects explicitly with less reliance on textures and other spurious image patterns, the reconstruction objective may unintentionally impose a barrier to success.

Limitations We note several limitations of this work to be addressed in future research. First, defining causal models of the image/corruption generating process is not trivial and we make no claim that our RobustCLEVR corruption variants model the "true" causal mechanisms or distributions for real-world image corruptions. We also acknowledge that the full space of possible causal model graphs, mechanisms, and distributions is intractable to evaluate. Nonetheless, we evaluate two contrasting causal models which are sufficient to successfully demonstrate that OC model performance is highly dependent on the SCM and underlying data distribution. We also did not explore causal dependencies between properties of the scene and the occurrence of corruptions (e.g., the presence of dark settings only for specific objects). However, our variants instead intend to capture a wide range of image distortions independent of the scene composition with the purpose of more broadly testing OC methods beyond the conventional IID case. Further, the corruptions in RobustCLEVR are applied late in the rendering pipeline which limits their overall realism. That said, CLEVR-like scenes are considerably simpler than real world data and the lack of robustness on RobustCLEVR images provides a useful check prior to testing on more complex scenes. Lastly, metrics are computed relative to ground truth image masks and clean images, yet in severe cases, corruptions will prevent OC methods from fully recovering the original objects/image. While this may make it difficult to estimate the true upper bound on performance, this does not prevent relative comparisons between models.

6. Conclusion

In light of recent advances in object-centric learning, we present the first benchmark dataset for evaluating robustness to image corruptions. To thoroughly test robustness, we adopt a causal model framework whereby assumptions about the corruption generating process can be explicitly implemented and compared. We evaluate a set of state-of-the-art OC methods on data generated from causal models encoding various assumptions about the corruption generating process. We find that OC models are not robust to corruptions and further demonstrate through our causal model framework that distributional assumptions matter when comparing model robustness. While our results indicate that OC models are not implicitly robust to a range of natural image corruptions, object-centric learning still holds great promise for achieving robust models in the future.

References

[1] Said Fahri Altindis, Yusuf Dalva, and Aysegul Dundar. Benchmarking the robustness of instance segmentation models. Sept. 2021. 3

[2] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering objects that can move. Mar. 2022. 3

[3] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. Jan. 2019. 3

[4] Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. Apr. 2021. 1

[5] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nat. Commun.*, 11(1):3673, July 2020. 3

[6] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. Dec. 2014. 3

[7] John Chen, Samarth Sinha, and Anastasios Kyrillidis. StackMix: A complementary mix algorithm. Nov. 2020. 1, 7

[8] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. *AAAI*, 33(01):3412–3420, July 2019. 3, 4

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. Sept. 2019. 1, 7

[10] Hao Ding, Jintan Zhang, Peter Kazanzides, Jie Ying Wu, and Mathias Unberath. CaRTS: Causality-Driven robot tool segmentation from vision and kinematics data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 387–398. Springer Nature Switzerland, 2022. 3

[11] Andrea Dittadi, Samuele Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and robustness implications in Object-Centric learning. July 2021. 1

[12] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. July 2020. 1

[13] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3D object detection to common corruptions in autonomous driving. 3

[14] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? Dec. 2021. 1, 3

[15] Yilun Du, Mit Kevin, Smith Mit, Joshua Tenenbaum, and Jiajun Wu. UNSUPERVISED DISCOVERY OF 3D PHYSICAL OBJECTS FROM VIDEO. 3

[16] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled Multi-Object representations. June 2021. 3, 4

[17] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-V2: Inferring unordered object representations without iterative refinement. Apr. 2021. 3, 4

[18] S M Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. Mar. 2016. 3

[19] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-Object representation learning with iterative variational inference. Mar. 2019. 3, 4

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. Mar. 2019. 1, 2, 3

[21] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. Dec. 2019. 1, 7

[22] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. May 2020. 3

[23] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. Oct. 2020. 3, 4

[24] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 2, 4, 5

[25] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3D common corruptions and data augmentation. Mar. 2022. 2

[26] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised multi-object segmentation by predicting probable motion patterns. Oct. 2022. 2, 3

[27] Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A Texture-Rich benchmark for unsupervised Multi-Object segmentation. Nov. 2021. 3, 4, 7

[28] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric learning from video. Nov. 2021. 3

[29] Pang Wei Koh, 1 Shiori Sagawa, 1 Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A Earnshaw, Imran S Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson 3 9 Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. 3

[30] Adam R Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential attend, infer, repeat: Generative modelling of moving objects. June 2018. 3

[31] Alfred Laugros, Alice Caplier, and Matthieu Ospici. Are adversarial robustness and common perturbation robustness independant attributes? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[32] Alfred Laugros, Alice Caplier, and Matthieu Ospici. Using synthetic corruptions to measure robustness to natural distribution shifts. July 2021. 3

[33] Jin-Ha Lee, Muhammad Zaigham Zaheer, Marcella Astrid, and Seung-Ik Lee. SmoothMix: a simple yet effective data augmentation to train robust classifiers, 2020. 1, 7

[34] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised Object-Oriented scene representation via spatial attention and decomposition. Jan. 2020. 3, 4

[35] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric learning with slot attention. June 2020. 3

[36] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. May 2016. 3

[37] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. Dec. 2020. 3

[38] Roman C Maron, Justin G Schlager, Sarah Haggenmüller, Christof von Kalle, Jochen S Utikal, Friedegund Meier, Frank F Gellrich, Sarah Hobelsberger, Axel Hauschild, Lars French, Lucie Heinzerling, Max Schlaak, Kamran Ghoreschi, Franz J Hilke, Gabriela Poch, Markus V Heppt, Carola Berking, Sebastian Haferkamp, Wiebke Sondermann, Dirk Schadendorf, Bastian Schilling, Matthias Goebeler, Eva Krieghoff-Henning, Achim Hekler, Stefan Fröhling, Daniel B Lipka, Jakob N Kather, and Titus J Brinker. A benchmark for neural network robustness in skin cancer classification. *Eur. J. Cancer*, 155:191–199, Sept. 2021. 2, 3

[39] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. July 2019. 2, 3

[40] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. Jan. 2022. 3

[41] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. PRIME: A few primitives can boost robustness to common corruptions. Dec. 2021. 1, 7

[42] Li Nanbo, Cian Eastwood, and Robert B Fisher. Learning Object-Centric representations of Multi-Object scenes from multiple views. Nov. 2021. 3

[43] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. Nov. 2020. 3

[44] Wei Qin, Hanwang Zhang, Richang Hong, Ee-Peng Lim, and Qianru Sun. Causal interventional training for image recognition. *IEEE Trans. Multimedia*, pages 1–1, 2021. 3

[45] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. Feb. 2021. 3

[46] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? June 2019. 1

[47] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised Object-Centric learning for complex and naturalistic videos. May 2022. 3

[48] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. July 2020. 1

[49] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization strategy to train strong classifiers with localizable features, 2019. 1, 7

[50] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an Object-Centric video representation beneficial for transfer? July 2022. 1

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. Oct. 2017. 1, 7

[52] Hua Zhang, Liqiang Xiao, Xiaochun Cao, and Hassan Foroosh. Multiple adverse weather conditions adaptation for object detection via causal intervention. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP, Apr. 2022. 3

[53] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. CausalAdv: Adversarial robustness through the lens of causality. June 2021. 3

[54] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. OOD-CV: A benchmark for robustness to Out-of-Distribution shifts of individual nuisances in natural images. Nov. 2021. 3