Causality-Driven Audits of Model Robustness

Nathan Drenkow[1,2] William Paul[1,2] Chris Ribaudo[1]
[1]The Johns Hopkins University Applied Physics Laboratory
Laurel, MD USA

Mathias Unberath[2]
[2]The Johns Hopkins University
Baltimore, MD USA

Abstract

*Robustness audits of deep neural networks (DNN) provide a means to uncover model sensitivities to the challenging real-world imaging conditions that significantly degrade DNN performance in-the-wild. Such conditions are often the result of multiple interacting factors inherent to the environment, sensor, or processing pipeline and may lead to complex image distortions that are not easily categorized. When robustness audits are limited to a set of isolated imaging **effects** or distortions, the results cannot be (easily) transferred to real-world conditions where image corruptions may be more complex or nuanced. To address this challenge, we present a new alternative robustness auditing method that uses causal inference to measure DNN sensitivities to the factors of the imaging process that **cause** complex distortions. Our approach uses causal models to explicitly encode assumptions about the domain-relevant factors and their interactions. Then, through extensive experiments on natural and rendered images across multiple vision tasks, we show that our approach reliably estimates causal effects of each factor on DNN performance using only observational domain data. These causal effects directly tie DNN sensitivities to observable properties of the imaging pipeline in the domain of interest towards reducing the risk of unexpected DNN failures when deployed in that domain.*

1. Introduction

A persistent challenge in the development and use of vision-based AI systems is dealing with the diversity of possible real-world operating conditions. Safety- and cost-critical applications require robust and reliable algorithms capable of maintaining their performance across diverse conditions seen during deployment. Recent studies [13, 20, 27, 52, 55] have shown that despite significant developments in deep learning methods, DNNs remain susceptible to performance degradation due to challenging natural imaging conditions. Robustness audits thus play a critical role in identifying the sensitivities of DNNs to these conditions before models are deployed in high-stakes applications. A compounding challenge is that the most demanding imaging conditions may arise as a result of multiple interacting factors related to the external environment (*e.g.*, lighting, weather, shadows), sensor (*e.g.*, exposure, ISO, F-stop), and processing pipeline (*e.g.* auto white balance, tone mapping, denoising).

Figure 1. Illustrative example of the "exposure triangle". Over- and under-exposure of the image may produce compounded effects with low contrast, low depth of field, and/or image noise.

As an illustrative example, consider the well-understood "exposure triangle" in natural photography where exposure time, aperture size (or F-stop), and ISO settings are adjusted to achieve the desired photographic appearance depending on lighting conditions. In limited illumination settings, captured images may contain combinations of effects such as limited depth of field (large aperture), blurring due to sensor/object motion (long exposure times), and higher noise (high ISO to account for low photon counts). Because it is not always possible to find an ideal compromise, all three distortions may occur simultaneously and in varying degrees of severity depending on the settings (Fig. 1). Furthermore, attempting to offset one effect (*e.g.*, exposure time) during imaging by adjusting one of the other settings (*e.g.*, ISO) may only result in a change in the nature of the distortion (*e.g.*, reduced motion blur but higher noise). Similarly, changes in environmental factors, such as time of day, weather, or location, will subsequently cause exposure, ISO, and F-stop settings to have varying impact on the overall image quality. Given this potential range of complexity in real-world conditions, we ask: **how can we identify the set of imaging factors that have the strongest effect on DNN robustness?**

In order to effectively audit the robustness of DNNs to

identify these factors, we need the following criteria. First, we need methods that can be grounded in domain knowledge and aligned with the types of imaging conditions expected in the target domain so that results obtained from an audit will be predictive of future performance in the targeted deployment conditions. Second, we need audits to produce interpretable and actionable insights. Audits should identify sensitivities to factors that developers can observe/control in training and deployment in order to improve model robustness. Finally, ideally it should be possible to conduct audits directly on complex real-world data whenever possible.

The current conventional form of robustness auditing uses the common corruptions framework [26], (Fig. 2a) which identifies a limited set of *effects* (*e.g.*, blur, noise, contrast) of the image generating process (IGP) and evaluates them independently (*i.e.*, no interactions). While this framework is capable of evaluating a range of conditions, it does not provide a sound theoretical framework for analyzing DNN sensitivities to complex conditions that occur when multiple imaging factors are compounded in real-world data (such as in the illustrative example above). Many of these evaluations are not grounded by specific domain knowledge and may produce findings that are difficult to translate to real-world settings. Similarly, by focusing on effects of the IGP, it is challenging to derive actionable insights for understanding or predicting how DNN performance may change as a result of changes to settings of the *causes* of the imaging conditions.

To address these limitations, we propose an alternative first-of-its-kind robustness auditing framework that analyzes DNN sensitivities with respect to the factors in the IGP that *cause* image distortions. Our audits are aligned with the imaging domain by using causal models to explicitly encode knowledge and assumptions about imaging factors and their interactions affecting image quality (such as the "exposure triangle" example in Fig. 2b). We can draw upon extensive knowledge of imaging pipelines (*e.g.*, natural [11, 33, 54] or medical CT [34, 59] domains) to identify relevant domain-specific imaging factors and compose them to form causal graphs for targeted domains of interest (Sec. 3). Using the tools of causal inference, our causality-driven robustness audits (CDRA) provide a theoretical framework for measuring changes in DNN performance *caused* by factors of the IGP while using real-world domain data containing a range of natural imaging conditions resulting from multiple interacting causal factors.

Contributions: We make the following contributions:

- We introduce our novel CDRA framework that enables analysis of model sensitivities to complex imaging domains comprised of multiple interacting factors (Sec. 3)
- We show that task DNN robustness to imaging factors depends heavily on properties of the domain encoded via causal models (Sec. 5.1)
- We show that CDRA is itself robust to errors in the specification of the causal model (Sec. 5.2)
- We demonstrate empirically that CDRA can be applied effectively to a wide range of possible domains and complex imaging conditions using observational data sampled directly from these domains (Sec. 5.1-5.3)

2. Related work

Robustness research in deep learning for computer vision largely falls into two categories: adversarial and natural/non-adversarial robustness. We focus here on *natural robustness* where there is no explicit attacker and performance of the DNN is measured against challenging, naturally-occurring conditions [17]. Robustness audits in this context are often framed around natural distribution shift [13, 27, 55], environmental/sensor shifts [2, 53], dataset differences [42], out-of-distribution (OOD) [58, 63], or common corruption robustness [26].

Conventional robustness audits: A majority of natural robustness auditing methods center around the common corruptions framework [26] which initially proposed 15 classes of corruptions each simulated *independently* at five discrete levels of severity. This approach was later expanded to other domains [14, 24] and tasks [1, 43, 44]. Related methods proposed a similar evaluation on a subset of common image perturbations [36, 37]. Robustness metrics like mean Corruption Error [26] or robustness score [36] are useful when distortions occur independently but are not suitable for measuring the impact of individual factors of the IGP on DNN performance in the presence of other interacting factors. The discrete and independent nature of these simulated corruptions and robustness evaluations prevents generalizing the results to real-world domains where multiple corruptions often co-occur, thus increasing the risk of unexpected DNN failures in deployment.

Causal inference for computer vision: Causal inference methods have shown promise for analyzing dataset bias [4, 32], shortcuts [16, 46], and size/composition [3, 57] and generating [15, 21] complex, large-scale datasets. We introduce a novel perspective here by focusing on specifying and using causal models of the IGP for *robustness evaluation* in contrast to methods that use causal models for representation learning [5, 48, 50], robust training [28, 56, 60, 62], causal discovery [41], and domain adaptation [61].

3. Methods

We now provide details on auditing task DNN robustness through the lens of causal inference.

3.1. Causal models of the image generating process

The basis for our robustness audit is the specification of a Graphical Causal Model (GCM) which explicitly codifies knowledge and/or assumptions about the image generating

(a) Analysis of *effects*: Common corruptions [26]

(b) Analysis of *causes*: **CDRA (ours)**

Figure 2. Causal graphs representing contrasting robustness auditing paradigms: (a) Common corruptions [26] and (b) **CDRA (ours)**. Each arrow represents a causal relationship and DNN performance is measured given labels (Y), predictions ($\hat{Y}$), and the desired metric ($M = f_M(Y, \hat{Y})$). The common corruptions approach focuses on a subset of possible *effects* (e.g., **B**lur, **C**ontrast, and **N**oise) but does not consider any interactions or co-occurrence between them (i.e., no arrows between {**B, C, N**}). In contrast, our CDRA approach enables analysis of DNN sensitivities to *causes* of image distortion (i.e., **L**ighting, **E**xposure, **F**-stop, **ISO**) while using images containing complex corruptions that result from multiple interacting factors that occur in real-world domains.

process of a domain $\mathcal{D}$. Formally, we start by specifying the set of primary imaging factors $\mathcal{V}$ in the domain (*e.g.* $\{L, E, F, ISO\} \subseteq \mathcal{V}$ in Fig. 2b). Each factor $V \in \mathcal{V}$ corresponds to an observable property of the environment (*e.g.*, time of day, weather, or location), sensor (*e.g.*, exposure, f-stop, ISO), or other aspect of the imaging pipeline (*e.g.*, white balance, compression) that impacts the formation and quality of the image X (*e.g.* Fig. 2b).

The GCM is specified in the form of a directed acyclic graph (DAG) $\mathcal{G}_\mathcal{D} = (\mathcal{V}, \mathcal{E})$ with variables $\mathcal{V}$ and directed edges $\mathcal{E}$ where edges encode pairwise causal relationships between these variables. The existence of directed edges $(U, V) \in \mathcal{E}$ between any pair of factors $U, V \in \mathcal{V}$ is tied explicitly to the assumptions and knowledge of the domain (*e.g.*, the relationship between how exposure time influences the ISO setting in natural photography). For many imaging domains, we may be able to fully specify the DAG given in-depth knowledge of the environment, sensor, and/or imaging process. While other domains may make it more difficult to specify the DAG, the graph can still be used as a targeted hypothesis about the imaging process. By stating this hypothesis explicitly, researchers and domain experts can refine the DAG over time through additional experimentation and deeper investigation of the imaging process.

Under the Markov assumption, a variable in the GCM is independent of its ancestors conditioned on its direct parents ($V \perp \mathbf{anc}(V) \mid \mathbf{pa}(V)$ where $\mathbf{anc}(V)$ and $\mathbf{pa}(V)$ are the ancestors and parents of V respectively). With this assumption, any factor is determined by $V = f_V(\mathbf{pa}(V), U_V)$ where $f_V(\cdot)$ is the causal mechanism and U_V is a latent term representing measurement noise or other exogenous stochasticity. *For our causal auditing method, we only require the DAG hypothesis and do not require any assumptions about the factor distributions (V, U_V) or causal mechanisms (f_V).*

3.2. Causality-driven robustness auditing

The objective of CDRA is to estimate the causal effects that individual factors ($V \in \mathcal{V}$) of the IGP have on task DNN performance while using image data collected under complex realistic conditions and accounting for potential interactions due to other factors in the DAG. We first choose a performance metric (*e.g.* correctness/accuracy) $M : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ for labels $\mathcal{Y}$. The prediction $\hat{Y} = f_{\hat{Y}}(X, U_X; \theta)$ is determined by the task DNN $f_{\hat{Y}}$ parameterized by θ and trained to approximate $P(Y|X)$. We assume that θ is fixed so that $f_{\hat{Y}}$ is deterministic and the latent noise term can be omitted. We augment our causal graph by adding edges from $X \rightarrow \{Y, \hat{Y}\}$ and $\{Y, \hat{Y}\} \rightarrow M$ (as in Fig. 2b). The metric calculation is given as $M = f_M(Y, \hat{Y}, U_M)$, and similarly, U_M can be dropped since M is typically deterministic.

Given the final DAG (including X and M), our goal is to estimate for each factor V the average causal effect (ACE) of that factor on M. Here ACE represents the effect that changing V has on M *while accounting for all other potential variable interactions that may impact* M. We formulate this as

$$ACE_M(V : v \rightarrow \tilde{v}) = \mathbb{E}[M|do(V = \tilde{v})] - \mathbb{E}[M|do(V = v)] \quad (1)$$

which measures the expected difference in M when we intervene (indicated by $do(\cdot)$) to set V to two different values $\tilde{v}, v$. The notion of intervention here is akin to removing the edges in the causal graph into V, setting the value of V to v, and resampling all remaining variable values according to the modified graph. This operation represents the *hypothetical* case where every image had been captured with $V = v$ (while all other variables were sampled according to the underlying distribution). The choice of $v, \tilde{v}$ allows for testing changes in performance over targeted factor ranges and future work could focus on conditional ACE or individual causal effects.

To estimate $ACE_M(V : v \rightarrow \tilde{v})$, we start from the causal estimand (*e.g.*, $\mathbb{E}[M|do(V = v)]$) and use the structure of the causal graph to determine whether we can convert this causal estimand to a statistical/observational one. This is the problem of *identifiability* in causal inference, and if we can reduce Eq. 1 to a purely statistical estimand, then V is *identifiable* and can be estimated directly from observational data. While our approach is not limited with respect to the

number of factors in the causal DAG, the DAG topology and potential presence of unobserved factors can impact identifiability. Because the sensor and/or imaging process for many applications and domains is well-understood, we can often make the assumption that all critical imaging factors affecting image quality are observable and accounted for in the GCM. Various techniques are available for identification and, in particular, backdoor/frontdoor adjustments [47] are used based on the structure of the causal graphs in our experiments. These techniques identify the particular subset of variables $\mathcal{W}$ that must be adjusted for in order to prevent confounding in the ACE estimates (see Appendix C for a detailed example).

Once we obtain a statistical estimand, many techniques are available for estimating $ACE_M(V: v \rightarrow \tilde{v})$ including traditional methods such as S-/T-/X-learners [35]. In our experiments, we use S-learners that are designed as models $\mu(w, v) = \mathbb{E}[M|W = w, V = v]$ where V is the targeted factor and W are the adjustment variables used to prevent biased ACE estimates. Then, we estimate $ACE_M(V: v \rightarrow \tilde{v})$ using the estimator $\hat{\mu}$:

$$\widehat{ACE}_M(V: v \rightarrow \tilde{v}) = \frac{1}{\mid \mathcal{D} \mid} \sum_{w \in \mathcal{D}} \hat{\mu}(w, \tilde{v}) - \hat{\mu}(w, v) \quad (2)$$

where $\hat{\mu}$ can be implemented with a variety of machine learning approaches. In general, the choice of technique for estimating ACE will depend on the nature of the variables and their relationships in the causal DAG.

3.3. Conducting CDRA

For CDRA, we estimate $ACE_M(V: v \rightarrow \tilde{v})$ for each $V \in \mathcal{V}$. When M is the typical correctness metric (*i.e.*, $\mathbb{1}[\hat{y} = y]$), we can interpret $ACE_M(V: v \rightarrow \tilde{v})$ as the change in accuracy when imaging factor V changes from v to $\tilde{v}$. In this case, we say that a DNN is robust to factor V when $ACE_M(V: v \rightarrow \tilde{v})$ is close to 0. Let v_0 be a nominal value for factor V such that any deviation away from nominal (i.e., $|\tilde{v} - v_0| > 0$) is likely to degrade the image quality, then we say a DNN is less robust for $ACE_M(V: v_0 \rightarrow \tilde{v}) < 0$ because the task DNN's performance is expected to decrease on average as a result of a change in V that degrades the imaging conditions.

It is important to note that because of the identifiability process (described above) and the observability of most key imaging factors, we can often obtain an estimate $ACE_M(V: v \rightarrow \tilde{v})$ purely from observational data without requiring the ability to directly modify or intervene on the actual image generation process. This means that we can collect a real-world evaluation dataset $\mathcal{D}$ under the expected imaging conditions of the domain and then estimate isolated causal effects of each imaging factor on DNN performance. Thus, we can evaluate DNNs on complex imaging conditions (with multiple corruptions) and still obtain estimates of the sensitivity of performance (via $ACE_M(V: v \rightarrow \tilde{v})$) to changes in specific observable imaging factors that *cause* those conditions.

4. Simulating complex image domains

To show experimentally that CDRA can expose DNN sensitivities in complex imaging domains, we require evaluation data with several properties. First, we need datasets where we can precisely control primary imaging factors and their interactions. Second, we need the ability to test our auditing approach across a diverse range of domains/conditions. Third, we need datasets that also still reflect the complexity of objects and scenes present in the real world. We see that previous datasets and benchmarks [26, 37, 44, 49] do allow for precisely controlling factors that lead to image degradation, but do not consider interactions amongst factors. Other approaches [15, 27] have introduced more complexity in simulating imaging domains, but have relied on scenes that lack aspects of real-world complexity. Lastly, while a number of public image benchmark datasets exist that cover a wide range of real-world scenes and conditions [6, 12, 29, 51], challenging conditions are often excluded and crucial metadata about the underlying imaging factors (typically encoded in EXIF tags) is removed (*e.g.*, due to privacy concerns, memory constraints, collection protocol).

To address these limitations, we generate a new set of synthetic datasets with complex imaging conditions that meet the necessary criteria above. We adapt the image corruptions from the ImageNet-C benchmark [26] and the causal model-based rendering from [15] to simulate a range of hypothetical domains whereby (1) we have full ground truth of the imaging process and underlying factors, (2) we can produce imaging conditions that go beyond the complexity of existing benchmarks and approach more natural, real-world settings, and (3) we can still audit the robustness of DNNs on real-world scenes. Our framework extends beyond the common corruptions approach by using *combinations of corruptions* as determined by a causal model to simulate more complex distortions than found in the conventional setting [26]. Using this data generation framework, we aim to assess the efficacy of our CDRA approach to uncover DNN sensitivities beyond what average performance metrics capture.

GCMs for domain-specific image generation: A primary benefit of the causal perspective on the IGP is that it enables a sparse factorization of complex distributions over factors that affect image quality. When we "invert" this property, we gain the ability to precisely control the generation of complex image distributions which we can use for evaluating the efficacy of CDRA. These complex distributions allow for simulating a wide range of imaging conditions whereby corruptions are combined to yield complex distortions.

To generate such distributions, we first define the domain $\mathcal{D}$ using the DAG $\mathcal{G}_{\mathcal{D}} = (\mathcal{V}, \mathcal{E})$. Underlying each

Table 1. Set of available corruptions as nodes in simulated GCMs.

Variable	Name	Variable	Name
GN	Gaussian noise	D	Defocus blur
N	Shot noise	C	Contrast
IN	Impulse noise	B	Brightness
S	Speckle noise	S	Saturate
G	Gaussian blur	P	Pixelate

domain is a joint distribution over all factors and images, $P_{\mathcal{G}}(\mathcal{V}, X) = \prod_{V \in \mathcal{V}} P(V|pa(V))$, which we can sample to generate evaluation datasets for our simulated domain. We use the GCM definition in conjunction with one of two corruption generation processes (*compositing* and *rendering*) to generate our evaluation datasets.

Compositing: In the compositing setting (adapted from [15]), we specify a set of image corruption functions (Table 1) to be *combined* according to the specified GCM DAG. Variables V in the GCM correspond to normalized severities for an associated set of corruption functions $\mathcal{C} = \{c_V \mid V \in \mathcal{V}\}$ where $c_V(x, V) = \tilde{x}$. We first sample the severity values for each corruption in the GCM (*i.e.* $V = f_V(pa(V), U_V)$ for all $V \in \mathcal{V}$). We then apply the corruption functions to the original image by following a topological ordering from $\mathcal{G}_{\mathcal{D}}$ and using the image output by the previous function in the ordering as input to the next corruption. The result is an image that has been distorted by multiple types of corruption (with differing levels of severity) and whose imaging conditions are more complex than if the image had only been modified by a single corruption in isolation (as in the common corruptions framework).

Rendering: In this work, we also extend beyond [15] by introducing a *rendering* approach which uses the causal model to directly guide physics-based rendering. Here, factors of the causal model correspond directly to settings of the rendering engine (*i.e.*, Blender [8]). These settings typically mirror measurements available when capturing image datasets in the real-world (*e.g.*, exposure, ISO, F-Number) and can be mapped directly to settings in Blender. To render corrupted images, we sample the factor values $\boldsymbol{v} \sim P_{\mathcal{G}}(\mathcal{V})$ and update the Blender settings directly with these values for a single scene. Then the physics-based `Cycles` engine renders the scene under the sampled conditions. The rendering process directly captures the full set of effects of the sampled factor values on the lighting, materials, scene geometry and dynamic properties of the scene.

5. Experiments

In the following experiments, we seek to show that CDRA allows us to accurately estimate how DNN performance changes due to changes in individual imaging factors in the GCM. We first examine the accuracy of CDRA in domains where we can simulate complex imaging conditions *and* perform interventions to compute ground truth $ACE_M(V : v \to \tilde{v})$. We then examine the effects of misspecification of the causal graph on estimating $ACE_M(V : v \to \tilde{v})$. Lastly, we apply CDRA to additional vision tasks and imaging domains to show that it generalizes and provides useful insights beyond image classification.

5.1. Auditing DNNs on diverse image domains

In order to demonstrate the generalizability of CDRA, we start by running audits over a set of complex imaging domains. For each simulated hypothetical domain, we evaluate a set of task models on all images in the domain. We then apply CDRA to compute $ACE_M(V : v \to \tilde{v})$, $\widehat{ACE}_M(V : v \to \tilde{v})$ for each factor/cause V in the domain GCM.

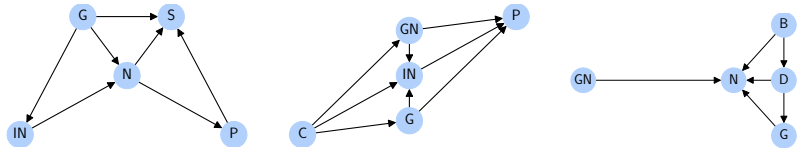
GCM sampling and image generation: We first specify a set of possible corruption functions $\mathcal{C}$ that act as proxies for observable variables $V \in \mathcal{V}$ in the domain simulated via the causal model. Given $\mathcal{C}$ (see Table 1), we generate a set of random GCMs by sampling a DAG of N variables $\mathcal{V}$ (with each $V \in \mathcal{V}$ associated with a corruption function $c_V \in \mathcal{C}$) and with probability $p((V_i, V_j) \in \mathcal{E}) = 0.5$ of a directed edge from $V_i \to V_j$ being formed between any pair of variables in the DAG. Also associated with each variable in the GCM is a set of conditional probability distributions (CPDs) that encode the relationship ($P(V \mid \mathbf{pa}(V))$) between the sampled value of the variable (V) and its parents ($\mathbf{pa}(V)$). We sample ten GCMs with $N = 5$ factors to allow for our imaging conditions to be sufficiently complex and in order to test whether CDRA can accurately estimate ACE values under such challenging conditions. We generate data for each simulated domain (GCM examples in Fig. 3; full set in Appendix A).

For each image in the ImageNet-Val dataset [12], we first sample the severity values for each $V \in \mathcal{V}$ from the distribution underlying the GCM and then apply the *multiple* corresponding corruptions following the compositing procedure outlined in Section 4 to the image. We limit severity values to $\{0, 1, 2\}$ (with 0 corresponding to $c_V(x, 0) = x$ and severity 2 included to allow for more variability in imaging conditions) in order to ensure that images corrupted with multiple functions will still be interpretable. For each simulated domain, we generate 50k samples with varying levels of corruption consistent with the GCM (see Appendix D for sample visualizations).

Ground truth ACE: To obtain ground truth $ACE_M(V : v \to \tilde{v})$ for comparison against the $\widehat{ACE}_M(V : v \to \tilde{v})$ estimates, we need to evaluate task DNNs on image data consistent with the interventional distributions in Eq.1 (*i.e.* $P(M|do(V = v)), P(M|do(V = \tilde{v}))$). For each GCM, we intervene by removing edges into V and setting V to the corresponding intervention value (*i.e.* $do(V = v)$). We resample the remaining variables according to the GCM's distribution and then render a new dataset with those sampled values. The ground truth $ACE_M(V : v \to \tilde{v})$ is computed

Table 2. **True** $ACE_{acc}(V: 0 \rightarrow 1)$ (%) **per variable** for a subset of GCMs (left side). While ACE is measured per factor, images are *corrupted by combinations of all factors* in the GCM. (right side) Common corruptions correspond to evaluation of individual corruptions in isolation (Fig. 2a). Values < 0 indicate that DNN accuracy decreases as corruption severity of V increases. ACE_{acc} close to 0 (or > 0) is an indicator of greater robustness. See App. A for true, estimated ACE_{acc} across all 10 GCMs.

GCM	0					1					2					Common corruptions framework				
DNN / Factor	G	IN	N	P	S	C	G	GN	IN	P	B	D	G	GN	N	G	GN	IN	N	P
ConvNext-B	-5.3	-4.5	-3.4	-6.9	-8.1	-22.5	-2.4	-4.9	-12.5	-2.4	-0.64	-8.3	-3.3	-1.8	-6.3	-8.2	-9.0	-12.4	-10.0	-9.5
ResNet50	-7.0	-7.0	-3.6	-5.3	-11.5	-27.2	-4.8	-5.9	-12.4	-3.9	0.04	-12.4	-4.7	-4.1	-13.6	-10.6	-16.7	-29.1	-19.0	-12.8
Swin-B	-4.8	-4.9	-4.0	-9.1	-7.3	-17.1	-3.8	-4.7	-9.9	-8.3	0.44	-9.1	-3.2	-1.8	-4.6	-9.3	-8.8	-10.8	-9.8	-9.4



(c) GCM 0 (d) GCM 1 (e) GCM 2

Figure 3. **DAGs for random GCMs (see also Table 2)** Images are rendered according to the corruption severities sampled from the distribution underlying each model.

Table 3. **Comparison of** Δ_{ACE} (%) averaged over all GCM factors. (Full table in Appendix B)

GCM	ConvNext-B	ResNet50	Swin-B
0	1.0	0.83	0.70
1	1.3	0.84	0.87
2	0.84	1.1	0.85
Mean (all GCMs)	0.91	0.76	0.79
Std (all GCMs)	0.79	0.74	0.72

as the difference in task DNN accuracy on the pair of interventional dataset variants. For our experiments, we calculate ground truth $ACE_{acc}(V: 0 \rightarrow 1)$ which measures the true change in task DNN accuracy when the corruption severity associated with V goes from 0 (no corruption) to 1 (presence of corruption). Our synthetic data generation process ensures the amount of image corruption is monotonic in V such that $ACE_{acc}(V: 0 \rightarrow 1)$ measures DNN sensitivities to the smallest increases in image corruption (as opposed to $\max_{a,b} ACE_M(V: a \rightarrow b)$ over all combinations of a, b). For comparison, we also compute ACE in the common corruptions framework where corruptions are applied individually and independently (corresponding to a GCM with no interactions between corruptions - see Fig. 2a).

Task DNNs: For all experiments, we evaluate task DNNs that cover a range of architectural design patterns, sizes, and levels of performance. In particular, we evaluate ResNet50 [25], ConvNext-B [40], and Swin-B Transformer [39] on each simulated GCM domain. All models are pretrained on ImageNet [12] and were not exposed to corrupted images during training, ensuring that DNN performance is not biased towards any GCM datasets.

CDRA setup: We conduct CDRA for each GCM and task model pair. We (initially) assume full knowledge of the GCM DAG for estimating causal effects. We use an S-Learner based on Random Forest regression for obtaining $\widehat{ACE}_M(V: v \rightarrow \tilde{v})$ since it is capable of capturing non-linear interactions in the data while limiting model bias in the ACE estimates. For each variable V and given $M := \mathbb{1}[\hat{Y} = Y]$ (*i.e.* correctness), we estimate the $\widehat{ACE}_M(V: v \rightarrow \tilde{v})$ and also compute the ground truth $ACE_M(V: v \rightarrow \tilde{v})$ as described above. We compute the ACE error as

$$\Delta_{ACE} = \left| \widehat{ACE}_M(V: v \rightarrow \tilde{v}) - ACE_M(V: v \rightarrow \tilde{v}) \right|.$$

Results: In Table 2, we see the per-node ground truth ACE_{acc} results computed for three (of ten) simulated domains associated with the corresponding GCM DAGs in Figure 3 (full results in Appendix A). The ACE_{acc} here measures how much a change in severity for a single corruption in the GCM can impact accuracy even when multiple types of corruption are present in the image (as commonly occurs in real-world domains). For comparison, the right side of Table 2 shows the ACE_{acc} when only a single corruption is applied in isolation (less common in real-world domains). First, we observe the general trend across GCMs and task DNNs that increasing the corruption severity of any node has a moderate negative effect on task DNN performance (average $ACE = -5\%$) and in some specific cases that impact is severe ($ACE = -22\%$). For a given imaging factor in each GCM, we observe ACE values vary across DNNs showing that each DNN is sensitive to different factors depending on the domain properties determined by the GCM. Similarly, we see that the magnitude of the ACE values for a given imaging factor may vary significantly from one domain GCM to another (*e.g.*, see G in GCMs 0, 1 and the single-factor case). Lastly, we observe differences in ACE_M across task DNNs and GCMs despite their similar average performance (Table A4).

We also compute $\widehat{ACE}_M(V: v \rightarrow \tilde{v})$ and measure Δ_{ACE} for each GCM individually using the ground truth ACE obtained as described in Section 5.1. The results are summarized in Table 3 and show that across all GCMs, the average error in $\widehat{ACE}$ is less than 1%.

Discussion: These results show that under complex imaging conditions, CDRA allows us to effectively estimate how individual factors of the image generating process influence the task DNNs performance. Whereas mean accuracy (the conventional robustness metric) is only able to summarize model performance at the dataset-level, CDRA provides

deeper insights into how individual factors of the domain more directly influence DNN behavior. The fact that average accuracy is similar yet there is variability in ACE values occurs across DNNs, imaging factors, and GCMs illustrates that observed DNN robustness depends heavily on the nature of the imaging domain and the types of conditions it produces. The low Δ_{ACE} values also show that we can accurately recover ACE_M from evaluations using observational data.

A primary and critical implication of these results is that CDRA enables analyzing fine-grained aspects of DNN behavior directly on domain imaging data collected under complex and diverse conditions. While previous work [26, 27, 37] has largely focused on computing mean performance metrics over images affected by isolated corruptions, CDRA allows for evaluation on image data containing multiple compounded corruptions that are more reflective of real-world imaging conditions. Differences in ACE_{acc} values between the factors in the sampled GCMs (Table 2 - left) and the single-factor cases (Tab. 2 - right) illustrate that evaluation results obtained via the common corruptions framework may not be predictive of DNN behavior in more complex imaging domains. This underscores the benefit of CDRA as it enables isolating sensitivities to imaging factors via ACE_{acc} while relaxing the data requirements to allow for direct evaluation on domain data with multiple causes of distortion.

5.2. Assessing sensitivity of CDRA to GCM DAG misspecification

The previous experiment demonstrated that CDRA could provide accurate, deep, and fine-grained insights into DNN performance. However, the results were contingent on the assumption that full knowledge of the GCM DAG is available. For many domains, this assumption is reasonable given that knowledge of the imaging process can be translated into an accurate specification of the DAG. However, other imaging domains may be complex, and misspecification of the causal DAG may result in errors in ACE estimates.

In this experiment, we test for edge-related DAG misspecifications as follows. For each GCM, we run CDRA using assumed DAGs that differ from the true domain DAG. To create DAG errors, we randomly sample up to $N_E \in \{1, 2, 4\}$ edges to add or delete from the DAG (where the number of edges added may be limited by the constraint that the graph must remain acyclic). We then run CDRA to estimate $\widehat{ACE}_M(V : v \rightarrow \tilde{v})$ for each V in the GCM using the misspecified DAG. We re-run this process five times with different selections of the missing/added edges. As in Sec. 5.1, we compute the error (Δ_{ACE}) to assess the impact of the DAG-related errors (Table B7). We compute a *residual error* by subtracting the baseline error for each DAG (see Table 3) from the new estimates to show how ACE estimation error is impacted by the DAG specification errors.

Table 4. **Effect of N_E DAG edge errors on ACE estimation.** Residual error is the deviation from the baseline estimation error (in %) when the DAG is correctly specified. Errors are averaged over all GCMs and corresponding factors. Close to 0 is best.

	Missing Edges								
DNN	ConvNext-B			ResNet50			Swin-B		
N_E	1	2	4	1	2	4	1	2	4
Mean	0.06	0.17	0.44	0.04	0.16	0.38	0.07	0.21	0.44
Std	0.50	0.79	1.3	0.45	0.73	1.0	0.44	0.78	1.2
	Additional Edges								
Mean	0.03	0.03	0.04	0.02	0.02	0.02	0.02	0.03	0.03
Std	0.11	0.13	0.13	0.10	0.13	0.14	0.11	0.13	0.11

Results: The primary results for testing robustness to DAG misspecification can be found in Table 4 (all detailed GCM results in Appendix B). We find that CDRA is robust to misspecifications in the GCM DAG with Δ_{ACE} changing by less than 1% on average from the baseline error even when the DAG differs by a substantial number of edges.

Discussion: Since under the Markov assumption from Sec. 3, the GCM DAG encodes assumed/known conditional independencies between variables, effects of edge errors may be relatively localized due to these independence assumptions. Furthermore, the identifiability process (Sec. 3) may include/exclude incorrect imaging factors in the process of estimating $ACE_M(V : v \rightarrow \tilde{v})$ when edge errors occur. However, the estimator itself (e.g., S-learner) may be robust to minor errors in these cases and still capable of producing accurate ACE estimates.

We can see that CDRA is able to provide detailed insights about DNN performance on complex image data even in the case that the assumed causal DAG includes edge-related errors. In general, we expect that identifying the primary variables in the DAG is tractable. While accurately identifying all relationships between these variables may be difficult, this experiment shows that CDRA can tolerate a small number of errors. Furthermore, since CDRA can be run efficiently (relative to the cost of evaluating DNNs on the image data), we can test several DAG hypotheses to determine whether ACE estimates are sensitive to edge errors when the true DAG is not fully known. In our experiment, we observe a low variance in Δ_{ACE} indicating that for multiple hypotheses for the same true DAG, the ACE estimates were still relatively close to the true value. These results were consistent across the full range of GCMs we evaluated against and indicate that CDRA generalizes well across diverse domains and potential errors.

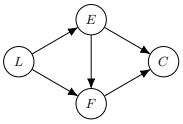
5.3. CDRA for additional vision tasks

Lastly, we have shown in Experiments 1, 2 that we can accurately estimate ACE (even in the presence of DAG specification errors), so we now use a similar experimental

Table 5. (a) ACE_{mIoU} by factor in the underlying DAG and average $mIoU$ by model for images generated according to the GCM in Figure 5b. Larger $|ACE_{mIoU}|$ indicates lower robustness of the OCL model to changes in the corresponding factor. (b) GCM DAG used to generate the CLEVR and MOVi-C evaluation datasets. {**L**: Lighting, **E**: Exposure, **F**: F-stop, **C**: Render cycles}

DNN / Factor	ACE_{mIoU} L	E	F	C	Dataset $mIoU$
GNM	12.7	1.2	2.3	-0.49	57.8
SPAIR	3.1	-0.34	2.8	1.6	57.5
IODINE	2.7	-0.16	-0.44	0.45	42.9
SPACE	12.4	1.7	-1.4	-1.2	36.1
GENESISv2	-6.7	0.48	1.3	1.3	27.4
EFFMORL	-8.9	0.99	1.1	1.7	17.6

(a)



(b)

setup show how CDRA can generate deeper robustness insights for additional vision tasks with GCMs that are better aligned with real-world settings. We use data *rendered* with Blender and GCMs specified for sampling the underlying physics-based rendering settings.

5.3.1. Object-Centric Learning

We first consider applying CDRA for an emerging class of vision algorithms for doing object-centric learning (OCL). We render corrupted versions of the CLEVR [31] dataset consisting of photo-realistic synthetically-rendered images of objects with varying colors, sizes, and materials. The causal model in Fig. 5b represents an implementation of the "exposure triangle" that we use for generating more realistic evaluation data (using the rendering method of Sec. 4). See App. E for more details.

We evaluate several baseline OCL algorithms including EffMORL [18], GENESISv2 [19], GNM [30], IODINE [22], SPACE [38], and SPAIR [10]. All models were trained on "clean" images similar to the original CLEVR benchmark. We measure task performance using mean Intersection over Union ($mIoU$) between the predicted segmentation masks and ground truth and use CDRA to estimate the ACE_{mIoU} for each factor of the GCM on the task performance of the DNN.

Results: Table 5a shows that causal effects expose sensitivities that are not obvious when looking at average $mIoU$ alone (last column of Table 5a). In particular, SPAIR and GNM are similar performing models, yet SPAIR appears to be more robust to changes in IGP factors as evidenced by the lower magnitude ACE values for factors of the GCM.

Discussion: These results underscore that both mean performance and ACE estimates are collectively necessary since $mIoU$ measures the performance of the DNN directly while ACE estimates sensitivity of the metric to changes in factor values. This is particularly evident in the case where IODINE exhibits low ACE values indicating high robustness to all factors of the IGP, yet its average $mIoU$ is lower than top DNNs. Similarly, SPAIR and GNM achieve higher $mIoU$, yet SPAIR appears to be more robust due to lower magnitude ACE values across most factors.

5.3.2. Optical Flow

We also apply CDRA to assess the robustness of optical flow (OF) methods using a GCM variant of the natural imaging "exposure triangle" (Fig. 5b). As in the OCL case, we find that even when top-performing baselines achieve similar average endpoint error (EPE), CDRA exposes sensitivities to IGP factors that differ between models. More detail on the experimental design and results are found in Appendix F.

6. Discussion and Conclusion

We present here a novel perspective on robustness auditing which uses causal inference to measure the sensitivity of DNN performance to *causes* of distortion in the image generating process. We find that even when average performance is similar between DNN models, the ACE estimates may differ measurably depending on the domain GCM and thus CDRA yields more granular insights into how model performance may change as a function of individual causes of image distortion.

Our approach is not free of limitations. First, our method is based on the ability to specify a plausible GCM DAG. This challenge is also faced by researchers applying causal inference in outside domains (*e.g.* medicine, economics, public health, ecology) where they are still able to generate deep insights into observational data even under uncertainty about the DAG [7, 9, 23, 45]. Deep knowledge of the environment, sensor, and imaging physics gives computer vision domain experts a distinct advantage in specifying accurate GCMs. Furthermore, while CDRA can be applied to real-world data without modification to the method itself, we were compelled to rely on simulated data due to the lack of public benchmarks containing sufficient metadata from the image capture process. In the near-term, our process for simulating hypothetical domains (Sec. 4) bridges the gap between the evaluation of task DNNs on isolated effects in the common corruptions framework and the complex conditions found in the real-world. By showing the efficacy of CDRA on simulated data in this work, our results highlight a gap in existing real-world benchmarks (with missing metadata) and encourage the collection of more comprehensive real-world datasets where new GCMs can be specified/refined in support of running CDRA on task DNNs of interest.

Overall, our CDRA method paves the way for enabling DNN robustness audits directly on domain-specific data containing complex, challenging imaging conditions. By measuring DNN sensitivities relative to causes in the IGP, the derived insights are actionable and can be used to develop new mitigation methods including targeted data collection/augmentation, architecture design, or GCM-driven adaptation.

References

[1] Said Fahri Altindis, Yusuf Dalva, and Aysegul Dundar. Benchmarking the robustness of instance segmentation models. 2021. 2

[2] Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. 2022. 2

[3] Bradley Butcher, Vincent S Huang, Christopher Robinson, Jeremy Reffin, Sema K Sgaier, Grace Charles, and Novi Quadrianto. Causal datasheet for datasets: An evaluation guide for Real-World data analysis and data collection design using bayesian networks. *Front Artif Intell*, 4:612551, 2021. 2

[4] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nat. Commun.*, 11(1):3673, 2020. 2

[5] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. 2014. 2

[6] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q H Truong, Chu The Chuong, and Curtis P Langlotz. CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats. 4

[7] Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. 110:12936–12941. 8

[8] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5

[9] Hannah E Correia, Laura E Dee, and Paul J Ferraro. Designing causal mediation analyses to quantify intermediary processes in ecology. 100:1512–1533. 8

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

[11] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annu Rev Vis Sci*, 7:571–604, 2021. 2

[12] 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. 4, 5, 6

[13] 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. 2020. 1, 2

[14] 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. *arXiv [cs.CV]*, 2023. 2

[15] Nathan Drenkow and Mathias Unberath. RobustCLEVR: A benchmark and framework for evaluating robustness in object-centric learning. 2023. 2, 4, 5

[16] Nathan Drenkow, Mitchell Pavlak, Keith Harrigian, Ayah Zirikly, Adarsh Subbaswamy, Mohammad Mehdi Farhangi, Nicholas Petrick, and Mathias Unberath. Detecting dataset bias in medical AI: A generalized and modality-agnostic auditing framework. 2

[17] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? 2021. 2

[18] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled Multi-Object representations. 2021. 8

[19] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-V2: Inferring unordered object representations without iterative refinement. 2021. 8

[20] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on ImageNet transfer to real-world datasets? 2023. 1

[21] Fabio Garcea, Lia Morra, and Fabrizio Lamberti. On the use of causal models to build better datasets. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1514–1519, 2021. 2

[22] 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. 2019. 8

[23] Gareth J Griffith, Tim T Morris, Matthew J Tudball, Anne Herbert, Giulia Mancano, Louise Pike, Gemma C Sharp, Jonathan A C Sterne, Kate Tilling, Luisa Zuccolo, Neil M Davies, and Gibran Hemani. Collider bias undermines our understanding of COVID-19 disease risk and severity. 31: 658–664. 8

[24] Haodong He, Jian Ding, and Gui-Song Xia. On the robustness of object detection models in aerial images. 2023. 2

[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] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. pages –, University of California, Berkeley, United StatesOregon State University, United States, 2019. International Conference on Learning Representations, ICLR. 2, 3, 4, 7

[27] Mark Ibrahim, Quentin Garrido, Ari Morcos, and Diane Bouchacourt. The robustness limits of SoTA vision models to natural variation. 1, 2, 4, 7

[28] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. 2020. 2

[29] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and Others. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597. 4

[30] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. 2020. 8

[31] 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. 8

[32] Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. No fair lunch: A causal perspective on dataset bias in machine learning for medical imaging. 2023. 2

[33] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *Computer Vision – ECCV 2016*, pages 429–444. Springer International Publishing, Cham, 2016. 2

[34] Oz Kilim, Alex Olar, Tamás Joó, Tamás Palicz, Péter Pollner, and István Csabai. Physical imaging parameter variation drives domain shift. *Sci. Rep.*, 12(1):21302, 2022. 2

[35] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. pages 4156–4165. 4

[36] 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

[37] Alfred Laugros, Alice Caplier, and Matthieu Ospici. Using synthetic corruptions to measure robustness to natural distribution shifts. 2021. 2, 4, 7

[38] 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. 2020. 8

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 10012–10022, 2021. 6

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6

[41] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. 2016. 2

[42] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. `http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf`. Accessed: 2023-4-4. 2

[43] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. COCO-O: A benchmark for object detectors under natural distribution shifts. 2023. 2

[44] 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. 2019. 2, 4

[45] Joost Oude Groeniger, Willem de Koster, and Jeroen van der Waal. Time-varying effects of screen media exposure in the relationship between socioeconomic background and childhood obesity. 31:578–586. 8

[46] Mitchell Pavlak, Nathan Drenkow, Nicholas Petrick, Mohammad Mehdi Farhangi, and Mathias Unberath. Data AUDIT: Identifying attribute utility- and Detectability-Induced bias in task models. 2023. 2

[47] Judea Pearl. *Causality*. Cambridge University Press, 2009. 4

[48] 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. 2

[49] Jenny Schmalfuss, Victor Oei, Lukas Mehl, Madlen Bartsch, Shashank Agnihotri, Margret Keuper, and Andrés Bruhn. RobustSpring: Benchmarking Robustness to Image Corruptions for Optical Flow, Scene Flow and Stereo. 4

[50] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. 2021. 2

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, P Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, R Kaczmarczyk, and J Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. abs/2210.08402. 4

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

[53] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 2

[54] R Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 2

[55] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. 2020. 1, 2

[56] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. 2021. 2

[57] Athanasios Vlontzos, Hadrien Reynaud, and Bernhard Kainz. Is more data all you need? a causal exploration. 2022. 2

[58] Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. Assaying Out-Of-Distribution generalization in transfer learning. 2022. 2

[59] Philip J Withers, Charles Bouman, Simone Carmignato, Veerle Cnudde, David Grimaldi, Charlotte K Hagen, Eric Maire, Marena Manley, Anton Du Plessis, and Stuart R Stock. X-ray computed tomography. *Nat. Rev. Methods Primers*, 1 (1):1–21, 2021. 2

[60] Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. 2023. 2

[61] 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, 2022. 2

[62] 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. 2021. 2

[63] 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. 2021. 2