# What makes domain generalization hard?

**Spandan Madan**
Harvard SEAS, Harvard-MIT CBMM
Cambridge, MA
spandan_madan@seas.harvard.edu

**Li You**
I2R, A*STAR
Singapore
liyou2001@gmail.com

**Mengmi Zhang**
CFAR and I2R, A*STAR, Harvard-MIT CBMM
Singapore
mengmi@i2r.a-star.edu.sg

**Hanspeter Pfister**
Harvard SEAS
Cambridge, MA
pfister@seas.harvard.edu

**Gabriel Kreiman**
Boston Children's Hospital, Harvard-MIT CBMM
Cambridge, MA
gabriel.kreiman@childrens.harvard.edu

## Abstract

While several methodologies have been proposed for the daunting task of domain generalization, understanding what makes this task challenging has received little attention. Here we present *SemanticDG* (Semantic Domain Generalization): a benchmark with 15 photo-realistic domains with the same geometry, scene layout and camera parameters as the popular 3D ScanNet dataset, but with controlled domain shifts in lighting, materials, and viewpoints. Using this benchmark, we investigate the impact of each of these semantic shifts on generalization independently. Visual recognition models easily generalize to novel lighting, but struggle with distribution shifts in materials and viewpoints. Inspired by human vision, we hypothesize that scene context can serve as a bridge to help models generalize across material and viewpoint domain shifts and propose a context-aware vision transformer along with a contrastive loss over material and viewpoint changes to address these domain shifts. Our approach (dubbed as *CDCNet*) outperforms existing domain generalization methods by over an 18% margin. As a critical benchmark, we also conduct psychophysics experiments and find that humans generalize equally well across lighting, materials and viewpoints. The benchmark and computational model introduced here help understand the challenges associated with generalization across domains and provide initial steps towards extrapolation to semantic distribution shifts. We include all data and source code in the supplement, and will make it publicly available upon publication.

## 1 Introduction

Domain generalization refers to the challenging setting where a model is trained on a set of related source domains and then evaluated on an out-of-distribution (OOD) target domain [1]. While several methodologies have been proposed for this task [2, 3, 4, 5, 6, 7, 8, 9], it remains unclear what aspects of domain shift make generalization challenging in the first place. This gap in understanding is in part because the distribution shift in existing benchmarks (for instance, *Photo* $\rightarrow$ *Cartoon*) can neither be quantified, nor disentangled into scene parameters which can be interpreted or

Corresponding Authors: spandan_madan@seas.harvard.edu, gabriel.kreiman@childrens.harvard.edu

controlled [10, 11, 12, 13]. To address this challenge, we present the *SemanticDG* benchmark. Our benchmark contains 15 different photo-realistic domains created by reconstructing 1,288 scenes from the popular ScanNet dataset [14] with the exact same geometry, layout and camera parameters, and controlled variations in scene lighting, object materials, and viewpoints. This benchmark is enabled by recent successes in inverse rendering [15] and the OpenRooms framework [16]. In Fig. 1, we show a real-world image from ScanNet, a paired photo-realistic image from the 3D scene reconstructed using inverse rendering and OpenRooms [16], and one image each with disentangled distribution shifts in materials, lighting and viewpoint respectively. Objects (e.g., table) are placed in context in these images to enable context-aware object recognition.

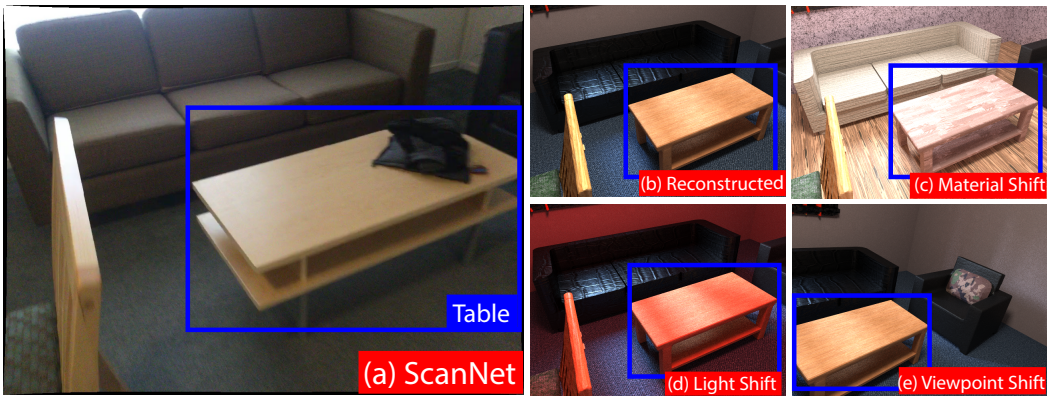



Figure 1: **SemanticDG benchmark:** (a) Real-world image from the ScanNet dataset. (b) Paired photo-realistic image created by reconstructing the real-world ScanNet image with matched object geometry, scene layout and camera parameters. (c) Image created by modifying the materials of the reconstructed image while holding viewpoint and light constant. (d) Image created by modifying the lighting of the reconstructed image while holding viewpoint and materials constant.(e) Image created by modifying the camera viewpoint of the reconstructed image while holding materials and light constant. Blue bounding boxes mark the target object (table) in all images.

In line with recent work using photo-realistic data to evaluate generalization behaviour [17, 18, 19, 20], we use our *SemanticDG* benchmark to isolate and quantify the impact of distribution shifts in lighting, viewpoint and material on generalization independently. Our findings reveal significant differences in the capability of networks to generalize across these transformations—while networks easily generalize to novel lighting, OOD materials and viewpoints present a major challenge.

To address this challenge, we propose a general-purpose solution to improve generalization across semantic domain shifts. Building on existing approaches for domain generalization that increase data diversity and enforce consistency across diverse views [3, 2, 4], we increase diversity by using multiple domains with disjoint sets of light settings, materials, and viewpoints. Furthermore, inspired by the role of scene context in human vision [21, 20, 22, 23, 24], we hypothesize that scene context can serve as a bridge to help generalize across semantic domain shifts, as contextual cues stay consistent between different semantic domains. Combining these ideas, we propose a Contextual Domain Contrastive Net (CDCNet). Our model consists of two separate streams to process the object and context independently before fusing them via cross-attention mechanism in a transformer decoder. To encourage CDCNet to learn robust and generic context information across domains, we apply supervised contrastive loss on the context-modulated object representations, which attracts the latent representations of the context where the target objects from the same object category are located, and repels the irrelevant context. Our proposed model outperforms state-of-the-art domain generalization methods by a **large margin of over 18%**, and domain adaptation methods by over **16%**.

As a critical benchmark, we also conduct psychophsyics experiments on *SemanticDG* to quantify the generalization capabilities of human vision. We find that humans can generalize equally well across materials, lights and viewpoints. Furthermore, humans still outperform computational models by a large margin in the real-world images, highlighting the challenges of domain generalization.

The key contributions in this study are:

- Releasing *SemanticDG*, the first domain generalization benchmark with controlled and disentangled semantic domain shifts.
- Quantifying what makes domain generalization hard—while lighting is easy, material and viewpoint shifts present a challenge for networks.
- Showing that, unlike models, humans generalize equally well across all three domain shifts.
- Introducing a context-aware transformer model with contrastive loss that outperforms both domain generalization and domain adaptation methods by a large margin, and learns to generalize from synthetic images to the real-world.

## 2 Related Work

### 2.1 Domain Generalization

Existing solutions for domain generalization often rely on increasing data diversity. This includes data augmentation techniques like MixUp [2], AugMix [3], RANDConv [4], style transfer based methods using AdaIN [25], and adversarial approaches to increase data diversity during training [5, 6, 7, 8, 9]. Note that unlike domain adaptation, domain generalization methods do not have access to unlabelled images from the test domain. Recent work by *Zhou et al.* provides a comprehensive review [26].

Beyond domain generalization, there is also significant work in generalization across transformations like 2D rotations and shifts [27, 28], and commonly occurring perturbations like blur or noise patterns [29, 30, 5, 31]. Recently, the community has also explored generalization to more global, real-world transformations using photo-realistic synthetic data, which include—3D rotations [32, 33], OOD category-viewpoint combinations [19] and incongruent scene context [20, 24]. In parallel, detecting OOD images has also received attention [34, 35, 36, 37, 38]. These works can broadly be divided into three categories—learning invariant representations [39, 40, 41, 42], causal representation learning to embed priors in the learning strategy [43, 44, 45], and custom optimization methods to enable generalization [46, 47, 48, 49].

Approaches relying on increasing data diversity are most closely related to our methodology. However, these do not use scene context or increase diversity using lighting, material and viewpoint variations like our proposed CDCNet.

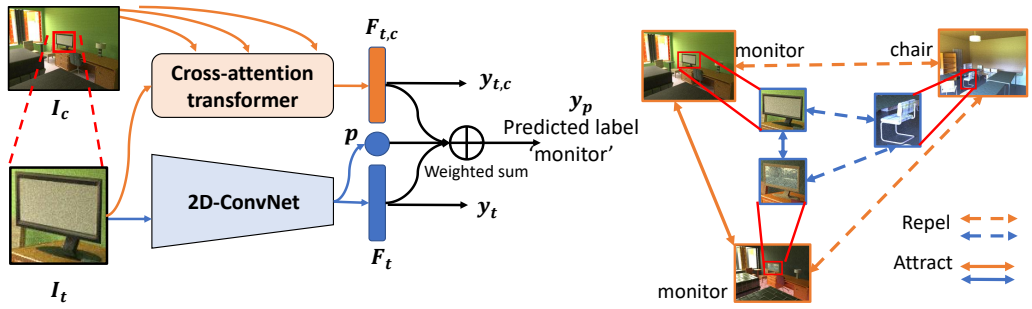
In recent years, several benchmarks have been proposed to study domain generalization, including—PACS [10], VLCS [11], Office-Home [12], and DomainNet [13], among others. These datasets contain a wide variety of domains including photos, cartoons, sketches, paintings, and other artistic renditions. However, a major challenge in these benchmarks is that the distribution shift cannot be quantified, controlled or disentangled into scene parameters. For instance in PACS [10], it is unclear how the *Photo* $\rightarrow$ *Cartoon* domain shift differs from a *Photo* $\rightarrow$ *Art* shift. Our proposed *SemanticDG* was designed specifically to address this challenge.

### 2.2 Scene context in human and computer vision

While object representations independent of scene context have become popular in the past few years, there is a long history of work studying the role scene context in human vision [21, 50, 51, 23, 52, 53, 24] and modelling contextual cues computationally [23, 54, 24, 20]. Recently, there have been some efforts to understand the impact of congruent and incongruent scene context on humans and computer vision [24, 20], and the recently proposed CRTNet [20] is designed to reason over scene context to better generalize to incongruent scene contexts. To the best of our knowledge, the role of context on domain generalization has not been explored before. Here we hypothesize that contextual reasoning can help with domain generalization.

## 3 Semantic Domain Generalization Benchmark (SemanticDG)

Quantifiable and interpretable domain shifts are at the heart of the *SemanticDG* benchmark. We recreated 1,288 real-world scenes from the 1,500 scenes available in ScanNet with the exact same 3D objects, scene layouts, class distributions and camera parameters (**Fig. 1a-b**). To achieve this, we relied on recent advances in inverse rendering [15] and the OpenRooms framework [16]. With these scenes, we created 15 different photo-realistic domains with 3 different types of domains




(a) Context-aware Feature Extraction (b) Contrastive Learning

Figure 2: **Architecture overview for Contextual Domain Contrastive Net (CDCNet)**. (a) Modular steps carried out by CDCNet in context-aware object recognition. CDCNet consists of 3 modules: feature extraction, integration of context and target information, and confidence-modulated classification. CDCNet takes the cropped target object $I_t$ and the entire context image $I_c$ as inputs and extracts their respective features. These feature maps are tokenized and information from the two streams is integrated over multiple cross-attention layers. CDCNet also estimates a confidence score $p$ for recognition using the target object features alone, which is used to modulate the contributions of $F_t$ and $F_{t,c}$ in the final weighted prediction $y_p$. (b) To help CDCNet learn generic representations across domains, we introduce contrastive learning on the context-modulated object representations $F_{t,c}$ in the embedding space. Target and context representations for objects of the same category are enforced to attract each other, while those from different categories are enforced to repel. Pairs for contrastive learning are generated using various material, lighting or viewpoint shifts (see **Sec. 4**).

shifts including—lighting shift, material shift, viewpoint shift (**Fig. 1c-e**). Each domain contains only one domain shift at a time. We rendered 19,800 images for each domain, which results in a total of 70k object instances from 13 indoor object categories overlapping with ScanNet (more details in Supplement). Across all domains, this amounts to roughly 300k images, and over 1 million object instances. Below, we explain how different domain shifts were created.

**Material shift domains:** We used 300 high quality, procedural materials from Adobe Stock including wood materials, fabrics, floor and wall tiles, and metals, among others. These were split into 6 sets of 50 materials each to create 6 different material domains, as shown in **Fig. 4(a)**. For each domain, its 50 materials were randomly assigned to scene objects to ensure high material diversity. One domain was held out for testing (OOD Materials), and the rest were used for training.

**Light shift domains:** As ScanNet scenes contain both indoor and outdoor lighting, we controlled outdoor lighting by using 300 different High Dynamic Range (HDR) environment maps from the Laval Outdoor HDR Dataset [55] and OpenRooms. These were split into 6 sets of 50 environment maps each (one set per domain). To create disjoint sets of indoor lighting, we split the HSV color space into 6 different chunks of disjoint hue values and sampled indoor light color and intensity from one chunk per domain. One domain was held out for testing (OOD Light), and the rest used for training.

**Viewpoint shift domains:** Controlling object viewpoints presents a challenge as objects like chairs and trash-cans are seen across a wide variety of azimuth angles (i.e., side vs front) across 3D scenes. Thus, to create disjoint viewpoint domains we chose to control the zenith angle by changing the height at which the camera is focusing. Due to difficulty in obtaining viable camera views with minimal occlusion, we created only four viewpoint train domains and one OOD Viewpoint domain.

**Geometry domain shift:** We created one additional test domain with deformed object meshes using a combination of three basic operations in blender—(1) randomly pushing or pulling object vertices, (2) applying a shear, and (3) mesh spherization by moving vertices towards the center. Sample images are available in the supplement. Only one test domain (OOD Geometry) was created with this shift. No models were trained with geometric deformations.

# 4 CDCNet: Contextual Domain Contrastive Net

A schematic of the proposed Contextual Domain Contrastive Net (CDCNet) is shown in **Fig 2**. We start with the Context-aware Transformer Network as the backbone [20] and introduce critical modification of contrastive learning described below to enable generalization across semantic shifts.

## 4.1 Feature Extraction in Context-aware Recognition using a Cross-attention Transformer

The context-aware recognition model in [20] achieved superior performance in in-context object recognition when the training and test data are from the same domain. Here, we used the same backbone and briefly introduce the network architecture below (see [20] for implementation details).

Given the training dataset $D = \{x_i, y_i\}_{i=1}^n$, CDCNet is presented with an image $x_i$ with multiple objects and the bounding box for a single target object location. $I_{i,t}$ is obtained by cropping the input image $x_i$ to the bounding box whereas $I_{i,c}$ covers the entire contextual area of the image $x_i$. $y_i$ is the ground truth class label for $I_{i,t}$. In this subsection, we focus on extracting context and target features in the embedding space and omit the index $i$ for simplicity. Inspired by the eccentricity dependence of human vision, CDCNet has one stream that processes only the target object ($I_t, 224 \times 224$), and a second stream devoted to the periphery ($I_c, 224 \times 224$) which processes the contextual area.

The context stream is a transformer decoder, taking $I_c$ as the query input and $I_t$ as the key and value inputs. The network integrates object and context information via hierarchical reasoning through a stack of cross-attention layers in the transformer, extracts context-integrated feature maps $F_{t,c}$ and predicts class label probabilities $y_{t,c}$ within $C$ classes.

A model that always relies on context can make mistakes under distribution shifts. Thus, to increase robustness, CDCNet makes a second prediction $y_t$, using only the target object information alone. A 2D CNN is used to extract feature maps $F_t$ from $I_t$, and estimates the confidence $p$ of this prediction $y_t$. Finally, CDCNet computes a confidence-weighted average of $y_t$ and $y_{t,c}$ to get the final prediction $y_p$. If the model makes a confident prediction with the object only, it overrules the context reasoning stage.

## 4.2 Supervised Contrastive Learning for Domain Generalization

Contrastive learning has benefited many applications in computer vision tasks (*e.g.*, [56, 57, 58, 59, 60]). However, all these approaches require sampling positive and negative pairs from real-world data. To curate positive and negative pairs, image and video augmentations operate in 2D image planes or spatial-temporal domains in videos. Here we introduce a simple and yet effective contrastive learning method on 3D synthetic data, resulting in promising generalization performance in real-world data.

Our contastive learning framework builds on top of the supervised contrastive learning loss [61]. Given the training dataset $D = \{x_i, y_i\}_{i=1}^n$, we randomly sample $N$ data and label pairs $\{x_k, y_k\}_{k=1}^N$. The corresponding batch pairs used for constrative learning consist of $2N$ pairs $\{\tilde{x}_l, \tilde{y}_l\}_{l=1}^{2N}$, where $\tilde{x}_{2k}$ and $\tilde{x}_{2k-1}$ are two views created with random semantic domain shifts of $x_k (k = 1, ..., N)$ and $\tilde{y}_{2k} = \tilde{y}_{2k-1} = \tilde{y}_k$. As elaborated in **Sec 3**, a domain shift is randomly selected from a set of *SemanticDG* domains specified during training. For example, if $x_k$ is from a material domain, $\tilde{x}_{2k}$ and $\tilde{x}_{2k-1}$ could be images from the same 3D scene but with different materials. For brevity, we refer to a set of $N$ samples as a batch and the set of $2N$ domain-shifted samples as their multiviewed batch.

Within a multiviewed batch, let $m \in M := \{1, ..., 2N\}$ be the index of an arbitrary domain shifted sample. Let $j(m)$ be the index of the other domain shifted samples originating from the same source samples belonging to the same object category, also known as the positive. Then $A(m) := M \backslash \{m\}$ refers to the rest of indices in $M$ except for $m$ itself. Hence, we can also define $P(m) := \{p \in A(m) : \tilde{y}_p = \tilde{y}_m\}$ as the collection of indices of all positives in the multiviewed batch distinct from $m$. $|P(m)|$ is the cardinality. The supervised contrastive learning loss takes the form:

$$L_{contrast} = \sum_{m \in M} L_m = \sum_{m \in M} \frac{-1}{|P(m)|} \sum_{p \in P(m)} \log \frac{\exp(z_m \cdot z_p / \tau)}{\sum_{a \in A(m)} \exp(z_m \cdot z_a / \tau)} \tag{1}$$

Here, $z_m$ refers to the context-dependent object features $F_{m,t,c}$ on $\tilde{x}_m$ after L2 normalization. The design motivation is to encourage CDCNet to attract the objects and their associated context from the same category and repel the objects and irrelevant context from different categories.

As previous works have demonstrated the essential role of context in object recognition [20, 24], contrastive learning on the context-modulated object representations enforces CDCNet to learn generic category-specific semantic representations across various domains. $\tau$ is a scalar temperature value which we empirically set to $0.1$.

Overall, CDCNet is jointly trained end-to-end with two types of loss functions: first, given any input $x_m$ consisting of image pairs $I_{m,c}$ and $I_{m,t}$, CDCNet learns to classify the target object using the cross-entropy loss with the ground truth label $y_m$; and second, contrastive learning is performed with features $F_{m,t,c}$ extracted from the context streams of CDCNet:

$$L = \alpha L_{contrast,c,t} + L_{classi,t} + L_{classi,p} + L_{classi,c,t} \tag{2}$$

Hyperparameter $\alpha$ is set to $0.5$ to balance the modulation effect of supervision from constrastive learning and classification loss.

# 5 Experimental Details

## 5.1 Baseline Architectures and Hardware details

We compared CDCNet against several baselines presented below. All models were trained on NVIDIA Tesla V100 16G GPUs. Our models converged in 30 hours, but domain generalization benchmarks (see below) were trained for 3 days till convergence. Optimal hyper-parameters for benchmarks were identified using random search, and all hyper-parameters are available in the supplement.

**Context-aware recognition models:** We include CRTNet [20] and Faster R-CNN [62]. CRTNet fuses object and contextual information with a cross-attention transformer to reason about the class label of the target object. We also use a modified Faster R-CNN [62] model to perform recognition by replacing the region proposal network with the ground truth location of the target object.

**2D feed-forward object recognition networks:** Previous works have tested popular object recognition models in generalization tests [63, 64]. We include the same popular architectures ranging from 2D-ConvNets to transformers: DenseNet [65], ResNet [66], MobileNetV2 [67], and ViT [68]. These models do not use context, and take the target object patch $I_t$ as input.

**Domain generalization and domain adaptation methods:** To evaluate the impact of context on generalization, we compare CDCNet to an array of state-of-the-art domain generalization methods (**Table 2**), as these methods do not use contextual information. They take $I_t$ alone as inputs. We also present comparisons with domain adaptation variants of these models. Note that these methods use unlabelled images from the test set during training, while our method does not.

## 5.2 Performance Evaluation

**Evaluation of Computational models.** *SemanticDG* comprises multiple training domains and one OOD test domain for each type of domain shift (light, viewpoint and material) (**Sec 3**). Top-1 classification accuracy is reported for computational models. We also include ScanNet images for testing generalization to the real world. See summary **Table 1** for experimental protocols.

**Human Behavioral Experiments.** Similar to the experiments with computer vision models, we evaluated generalization capabilities of humans on the *SemanticDG* benchmark using Amazon Mechanical Turk (MTurk). In each trial, subjects were presented with a fixation cross (500 ms), followed by a bounding box indicating the target object location (1000 ms). The image was shown for 200 ms. After image offset, subjects typed one word to identify the target object. See Supplement for the experiment schematics. We recruited 20 subjects per experiment, yielding 4,160 trials. Since humans have been regularly exposed to the object categories from *SemanticDG* in the real world, we removed the training phase for human experiments and directly tested humans with the same test set as computational models. To avoid choice biases and potential memory effects, we followed guidelines discussed in existing literature [20]. Additional details are provided in the Supplement.

# 6 Results

Below we present our findings on generalization capabilities of visual recognition models and humans across semantic domain shifts. We focus on five questions—(1) What makes domain generalization

| Section | Train Domains | Test Domains |
|---|---|---|
| Sec. 6.1 | Light 1D | OOD Light |
| | Material 1D | OOD Material |
| | Viewpoint 1D | OOD Viewpoint |
| Sec. 6.2 | Light 1D, 2D, 3D, 4D | OOD Light |
| | Material 1D, 2D, 3D, 4D | OOD Material |
| | Viewpoint 1D, 2D, 3D, 4D | OOD Viewpoint |
| Sec. 6.3 | Material 1D, 5D | OOD Material |
| Sec. 6.4 | Light 5D | OOD Light, OOD Mat, OOD Viewpoint, OOD Geometry, ScanNet |
| | Material 5D | OOD Light, OOD Mat, OOD Viewpoint, OOD Geometry, ScanNet |
| | Viewpoint 4D | OOD Light, OOD Mat, OOD Viewpoint, OOD Geometry, ScanNet |
| | All Domains | OOD Light, OOD Mat, OOD Viewpoint, OOD Geometry, ScanNet |

Table 1: **Summary of experimental protocols. Sec. 6.1**: For each type of domain shift (light, material, viewpoint), a model is trained with one training domain (1D) and tested on the corresponding OOD domain of the same type. **Sec. 6.2**: For each shift type, 4 models are trained with 1, 2, 3, and 4 training domains respectively, and all models are tested on the corresponding OOD domain. **Sec. 6.3**: To compare CDCNet with domain generalization and domain adaptation baselines, for each architecture we train a model with one material domain, and another with five material domains from *SemanticDG*. All models are then evaluated on the OOD Material domain. **Sec. 6.4**: Models trained on all available domains for one type are evaluated on OOD domains of different shift types.

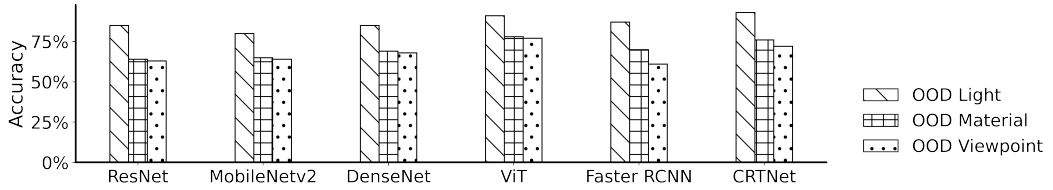



Figure 3: Generalization under light, material and viewpoint distribution shifts. Chance level = 7.6%. All architectures consistently generalize to new lighting easily, while material and viewpoint domain shifts present a challenge.

hard?, (2) How much does data diversity help?, (3) Do contextual cues help?, (4) Do models trained for one type of domain shift generalize to other types of shifts?, and (5) How well do humans generalize across these shifts?

### 6.1 Models generalize better across light as compared to material and viewpoint shifts

**Fig. 3** reports the impact of different semantic domain shifts on a wide variety of architectures including popular ConvNets, vision transformer, CRTNet, and FasterRCNN modified for recognition as described in **Sec. 5.1**. For each domain shift type (light, material, and viewpoint), one model was trained per architecture using one single training domain. These models were then evaluated on the corresponding OOD domain with the same type of domain shift. There is a consistent trend across all architectures—generalizing to new lighting is easy, while networks struggle more with both material and viewpoint domain shifts. Despite training with only 20% of light settings (HSV space for indoor light, and environment maps for outdoor light), models generalize well to OOD lighting. However, there is significant room for improvement in material and viewpoint shifts.

### 6.2 Data diversity improves genealization

In **Fig. 4(b)** we present the impact of increasing data diversity on generalization across different domain shift types. For each shift type, we trained 4 different models with the CDCNet architecture using 1,2,3 and 4 training domains respectively. These models were then evaluated on the corresponding OOD domain with the same type of shift.

Increasing data diversity helps increase performance across all three semantic domain shifts. Performance on OOD lighting quickly plateaus, with just two domains being sufficient. A key take-away here is that increasing material diversity can significantly help generalization to unseen materials, as

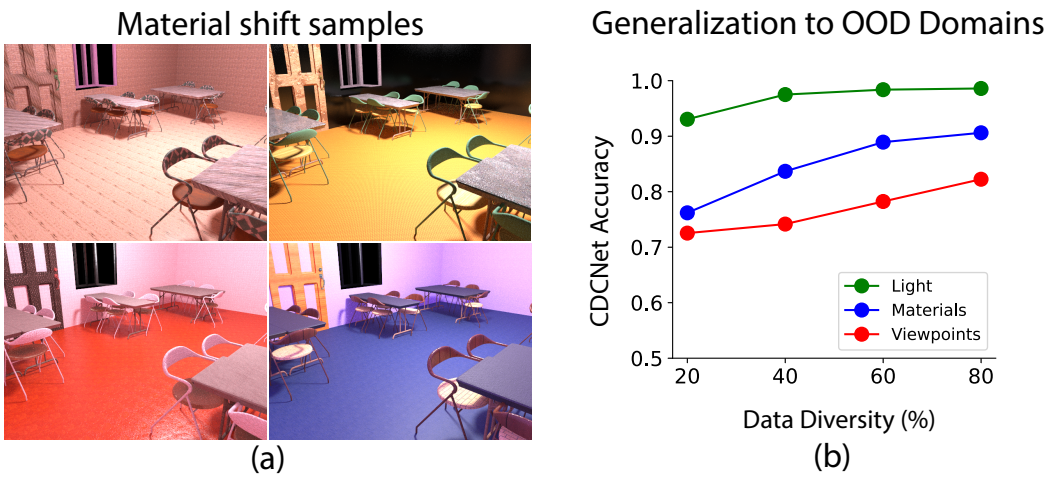



Figure 4: **Role of data diversity.** (a) Sample images from four material shift domains. 3D Object geometry, scene layout, lighting, camera viewpoint are the same for all images. Only the materials assigned to objects are changed. (b) Performance of CDCNet as data diversity is increased. CDCNet easily generalizes to OOD lighting, and can also generalize to OOD materials with high data diversity. However, there is significant room for improvement in generalizing to OOD viewpoints.

the accuracy jumps from 0.76 to 0.94. Thus, with appropriate material diversity, CDCNet is able to solve generalization to OOD materials. Note that procedural materials are highly diverse, representing complex surfaces like metals, rocks, fabrics, and even account for photo-metric attributes like wet vs dry rock, or rusted vs shiny metal. Thus, the increase from 50 to 200 materials during training represents a small fraction of all possible materials. This suggests that material changes, like lighting, are not the greatest challenge for domain generalization. The greatest challenge to generalization comes from OOD viewpoints. While increasing diversity can improve performance, there is still significant room for improvement.

### 6.3 Contextual cues improve generalization

In **Table. 2**, we report the results for generalization to OOD Materials for CDCNet along with a suite of domain generalization (DG) and domain adaptation (DA) methods. We also present results for the modified FasterRCNN model. For each architecture, we train two models—one trained with a single training domain, and another with five training domains. Note that domain adaptation methods had access to unlabelled images from the OOD Materials domain, while CDCNet, FasterRCNN and domain generalization methods did not. As data diversity increases to 5 domains, our method relying on the contrastive loss over contextual cues outperforms both domain generalization and domain adaptation models by a large margin of 18% and 16%, respectively. Thus, even with no information about the OOD test set, CDCNet outperforms specialized methods using unlabelled images from the test domain. This strongly suggests that scene context has a great impact on generalization across semantic domain shifts. The benefits of scene context significantly outweigh the benefits of using unlabelled test domain images.

### 6.4 Models generalize to real-world data

We assessed how well models trained for one semantic shift perform on other semantic shifts. For each semantic domain shift, we test CDCNet trained on all training domains of one type on OOD domains of different types. For instance, the model trained on 5 material domains was tested on OOD Light, OOD Viewpoint and OOD Geometry from *SemanticDG*. Models do not generalize across other semantic domains as well as they generalize within the domain shift they were trained for. Furthermore, OOD geometry causes a large drop in performance, suggesting this to be another major challenge for domain generalization.

As a real-world test for our benchmark, we also test how well our models trained on *SemanticDG* generalize to real images from the ScanNet dataset. For this, we collected roughly 8,000 object

| Task | Num Doms | AND Mask [69] | CAD [70] | COR AL [71] | ERM [72] | IRM [39] | MTL [73] | Self Reg [60] | VREx [74] | Faster RCNN [62] | CDCNet (ours) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DG | 1 | 0.748 | 0.756 | 0.756 | 0.748 | 0.749 | 0.748 | 0.760 | 0.736 | 0.7 | **0.762** |
| | 5 | 0.745 | 0.749 | 0.749 | 0.749 | 0.742 | 0.744 | 0.744 | 0.747 | 0.72 | **0.937** |
| DA | 1 | 0.755 | 0.766 | 0.783 | 0.783 | 0.776 | 0.783 | 0.767 | 0.780 | N/A | N/A |
| | 5 | 0.755 | 0.766 | 0.783 | 0.783 | 0.776 | 0.783 | 0.767 | 0.780 | N/A | N/A |

Table 2: **Role of contextual cues in generalization to OOD Materials.** We compare CDCNet to several domain generalization (DG) and domain adaptation (DA) methods that do not use contextual cues. We also report numbers for the modified FasterRCNN architecture which has access to scene context. CDCNet trained on five domains beats DG methods by over 18% and DA methods by over 16%, suggesting that contextual cues are substantially helpful in generalization across domain shifts.

| Train Domain | OOD Light | OOD Material | OOD Viewpoint | OOD Geometry | ScanNet |
|---|---|---|---|---|---|
| Light 5D | - | 0.79 | 0.76 | 0.61 | 0.46 |
| Material 5D | 0.96 | - | 0.84 | 0.62 | 0.46 |
| Viewpoint 4D | 0.73 | 0.73 | - | 0.47 | 0.48 |
| All domains | - | - | - | 0.58 | 0.48 |
| Humans | 0.67 | 0.65 | 0.67 | - | 0.66 |

Table 3: Generalization performance of our CDCNet and humans beyond trained domain type.

instances from ScanNet spanning our 13 object categories. Our models come close to the human upper bound of 0.66 without ever seeing real world images and despite the fact that ScanNet presents a domain shift in all three semantic domains—lighting, materials and viewpoints.

### 6.5 Humans can generalize across domains in the SemanticDG benchmark

Humans generalize equally well across the OOD light, material, and viewpoint domains (**Table 3**). This is in contrast with computational models (**Sec. 6.1**). Furthermore, humans achieve the same accuracy of 0.66 on real-world ScanNet images, and thus, perform equally well on real-world and domain-shifted images. This observation is again in contrast with computational models, as they all reveal a drop in accuracy beyond the training domain. Human accuracy also outperforms the best computational model on generalizing to ScanNet images. These findings suggest that humans are still superior at domain generalization. Despite these findings, we emphasize that our model CDCNet achieves significant success in closing the domain generalization gap compared to other baselines, and also generalizes well to real-world images without being trained with real images.

A fair comparison between humans and models is challenging. While models are trained and tested on images constructed with strict controls over the distribution shifts, the real world distribution humans learned from is far richer and more complex to pin down. We discuss the key differences between human and computational models in depth in the supplement.

## 7 Discussion

In the path towards general intelligence, we aspire to build algorithms that can readily extrapolate to novel settings. Such an algorithm should be capable of recognizing objects across any color or shape, scene layout, during day or night, in photographs, videos and even cartoon drawings. While we are far from such extreme domain generalization today, understanding and characterizing the capabilities of existing algorithms is essential to guide the AI community to prioritize the right set of problems. Here we introduce SemanticDG, which enables rigorous and quantitative analysis of the limits of domain generalization methods for recognition of objects embedded in complex scenes.

Through controlled analyses, we show that existing algorithms tend to generalize across light changes better than material changes, and that both of these semantic shifts are considerably easier than viewpoint changes. Pairing synthetic images with well-controlled real photographs with the same objects and layout, we compare human and computer vision in their capability to generalize across

these distribution shifts, and to recognize objects in real-world images. These results support the general conclusion that exposure to a large array of relevant domains leads to better generalization.

Interestingly, we find several key differences between human and computer vision, which merit further analysis—humans generalize equally well across all domain-shifted and real-world images, and serve as an aspirational upper bound for computational methods.

The datasets introduced here, including the paired synthetic and real images, and the human benchmark metrics provide a strong testbed for further algorithmic development to achieve better domain generalization. This is demonstrated in our proposed CDCNet, which brings together ideas from neuroscience, computer graphics and computer vision in a two-stream model that incorporates object and contextual information, and outperforms existing models by a large margin.

## References


[1] Alexander Robey, George Pappas, and Hamed Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[3] 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. *arXiv preprint arXiv:1912.02781*, 2019.

[4] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020.

[5] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[6] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[7] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[8] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[9] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

[11] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[12] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

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

[14] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[15] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020.

[16] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020.

[17] Guillaume Leclerc, Hadi Salman, Andrew Ilyas, Sai Vemprala, Logan Engstrom, Vibhav Vineet, Kai Xiao, Pengchuan Zhang, Shibani Santurkar, Greg Yang, et al. 3db: A framework for debugging computer vision models. *arXiv preprint arXiv:2106.03805*, 2021.

[18] Spandan Madan, Tomotake Sasaki, Tzu-Mao Li, Xavier Boix, and Hanspeter Pfister. Small in-distribution changes in 3d perspective and lighting fool both cnns and transformers. *arXiv preprint arXiv:2106.16198*, 2021.

[19] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146–153, 2022.

[20] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021.

[21] Eelke Spaak, Marius V. Peelen, and Floris P. de Lange. Scene context impairs perception of semantically congruent objects. *Psychological Science*, 33(2):299–313, 2022. PMID: 35020519.

[22] Peter De Graef, Dominie Christiaens, and Géry d'Ydewalle. Perceptual effects of scene context on object identification. *Psychological research*, 52(4):317–329, 1990.

[23] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[24] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020.

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[26] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[27] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

[28] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021.

[29] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[30] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[32] Avi Cooper, Xavier Boix, Daniel Harari, Spandan Madan, Hanspeter Pfister, Tomotake Sasaki, and Pawan Sinha. To which out-of-distribution object orientations are dnns capable of generalizing? *arXiv preprint arXiv:2109.13445*, 2021.

[33] Akira Sakai, Taro Sunagawa, Spandan Madan, Kanata Suzuki, Takashi Katoh, Hiromichi Kobashi, Hanspeter Pfister, Pawan Sinha, Xavier Boix, and Tomotake Sasaki. Three approaches to facilitate dnn generalization to objects in out-of-distribution orientations and illuminations: late-stopping, tuning batch normalization and invariance loss. *arXiv preprint arXiv:2111.00131*, 2021.

[34] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

[35] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[36] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[37] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[38] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[40] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

[41] Chuanlong Xie, Haotian Ye, Fei Chen, Yue Liu, Rui Sun, and Zhenguo Li. Risk variance penalization. *arXiv preprint arXiv:2006.07544*, 2020.

[42] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

[43] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

[44] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

[45] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[46] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[47] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[48] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

[49] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

[50] Michelle R Greene. Statistics of high-level scene context. *Frontiers in psychology*, 4:777, 2013.

[51] Peter De Graef. Scene-context effects and models of real-world perception. In *Eye movements and visual cognition*, pages 243–259. Springer, 1992.

[52] John M Henderson. Object identification in context: the visual processing of natural scenes. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3):319, 1992.

[53] Katherine Mary Mathis. *Does scene context automatically influence object recognition? Evidence from an interference task*. State University of New York at Albany, 1998.

[54] Jordan A Taylor and Richard B Ivry. Context-dependent generalization. *Frontiers in Human Neuroscience*, 7:171, 2013.

[55] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6927–6935, 2019.

[56] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[57] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[58] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[59] Mengmi Zhang, Tao Wang, Joo Hwee Lim, Gabriel Kreiman, and Jiashi Feng. Variational prototype replays for continual learning. *arXiv preprint arXiv:1905.09447*, 2019.

[60] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[61] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[63] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[64] Aidan Boyd, Kevin W Bowyer, and Adam Czajka. Human-aided saliency maps improve generalization of deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2735–2744, 2022.

[65] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[66] 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.

[67] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[68] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[69] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

[70] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

[71] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, abs/1607.01719, 2016.

[72] Ramakrishna Vedantam, David Lopez-Paz, and David J Schwab. An empirical investigation of domain generalization with empirical risk minimizers. *Advances in Neural Information Processing Systems*, 34, 2021.

[73] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

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

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

[76] Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018.

## Section S1 Sample images from all semantic shift domains

Below we present additional images from the *SemanticDG* benchmark. Each figure shows change in one scene parameter, while holding all others constant. Fig. S1 shows five different scenes from two training domains with a material shift. Similarly, in Fig. S2 we show images from two different light domains. Note that the first three rows in Fig S2 show different indoor lighting conditions controlled using indoor light color and intensity sampled from disjoint chunks of the HSV space. The last two rows show different outdoor lighting settings created by changing the environment maps. Similarly, Fig. S3 shows viewpoint shifted domains, and Fig. S4 shows the out-of-distribution (OOD) geometry domain created by deforming 3D meshes. As mentioned in the main paper, this domain was only used for testing—no models were trained on deformed geometry.

Figure S1: Example images from material domain shift. All other parameters held constant.

Figure S2: Example images from lighting domain shift. All other parameters held constant.

Figure S3: Example images from viewpoint domain shift. All other parameters held constant.

Figure S4: Example images from geometric deformation domain shift. All other parameters held constant.

| Task | Num Doms | AND Mask [69] | CAD [70] | COR AL [71] | ERM [72] | IRM [39] | MTL [73] | Self Reg [60] | VREx [74] | CDCNet (ours) |
|---|---|---|---|---|---|---|---|---|---|---|
| DG | 4 | 0.773 | 0.793 | 0.799 | 0.801 | 0.800 | 0.766 | 0.803 | 0.790 | **0.83** |
| DA | 4 | 0.751 | 0.786 | 0.772 | 0.774 | 0.768 | 0.800 | 0.792 | 0.779 | N/A |

Table S4: **Role of contextual cues in generalization to OOD Viewpoints.** We compare CDCNet to several domain generalization (DG) and domain adaptation (DA) methods that do not use contextual cues. CDCNet trained on four domains beats both DA and DG methods, suggesting that contextual cues are substantially helpful in generalization across domain shifts. The best results are bolded.

| Shift type | Num Doms | No context | Only context | Empty baseline | Two-stream Contrastive | No contrastive loss | Full model |
|---|---|---|---|---|---|---|---|
| Material Shift | 5 | 0.5 | 0.33 | 0.08 | 0.92 | 0.89 | **0.94** |

Table S5: **Impact of ablating and modifying components of CDCNet.**

## Section S2 Creating ScanNet test set for real-world experiments

To create the real-world test dataset, we sampled images from the ScanNet dataset [14]. ScanNet contains 1,500 3D scanned scenes, for which the frames of the captured video can be extracted. As the video is continuous, nearby frames are highly similar. To create a test set, we sampled one frame for every 100 frames in the video. Further, we only retained images which contained multiple overlapping object categories with OpenRooms. Using this method, we sampled 2-3 images per scene from 1,288 scenes, which resulted in roughly 2,900 test images with multiple object instances. In total, this set included 7,800 object instances.

## Section S3 Baseline models for generalization across viewpoint shifts

In table S4 we present results comparing our model CDCNet to existing benchmarks for domain adaptation and domain generalization. All architectures were trained on all available (4) viewpoint domains, and then tested on the held out OOD Viewpoint domain. As can be seen, CDCNet beats all domain generalization benchmarks.

Furthermore, the domain adaptation baselines reported here use unlabelled images from the testing domain, unlike CDCNet and other domain generalization methods. Despite not using this additional information, CDCNet outperforms domain adaptation baselines. This suggests that contextual cues are highly useful in generalizing across domain shifts, and can even be more impactful than using unlabelled images from the test distribution.

## Section S4 Ablations and Modifications for CDCNet

To quantify the contributions of different aspects of our model on generalization, we conducted several ablation studies with the material shift domain. We also try a modification where the contrastive loss is applied to both streams of CDCNet. All models were trained with five material shift domains, and then tested on the held out OOD Materials domain. These model variations are:

**No context:** To pass only target object information and no context, we set the context region (i.e., image region outside the target object) to all zeros during training.

**Only context:** To pass only context information, we set the target object region to all zeros, passing only the contextual information around it for training.

**Empty Baseline:** To get a good estimate of random chance, we pass all zeros into the network during training, while using the correct labels. That is, both context and target objects were set to zero.

**No contrastive loss:** We trained the model with no contrastive loss over different material, lighting and viewpoint domain shifts of the original image.

**Two-stream contrastive:** Compared with our full CDCNet, an additional supervised contrastive loss is applied here. Contrastive loss is calculated both for features $F_{m,t,c}$ and $Fm, t$ streams of CDCNet. Thus, CDCNet is jointly trained with the following losses:

$$L = \alpha(L_{contrast,c,t} + L_{contrast,t}) + L_{classi,t} + L_{classi,p} + L_{classi,c,t} \quad (3)$$

Hyperparameter $\alpha$ is set to $0.5$ to balance the modulation effect of supervision from constrastive learning and classification loss.

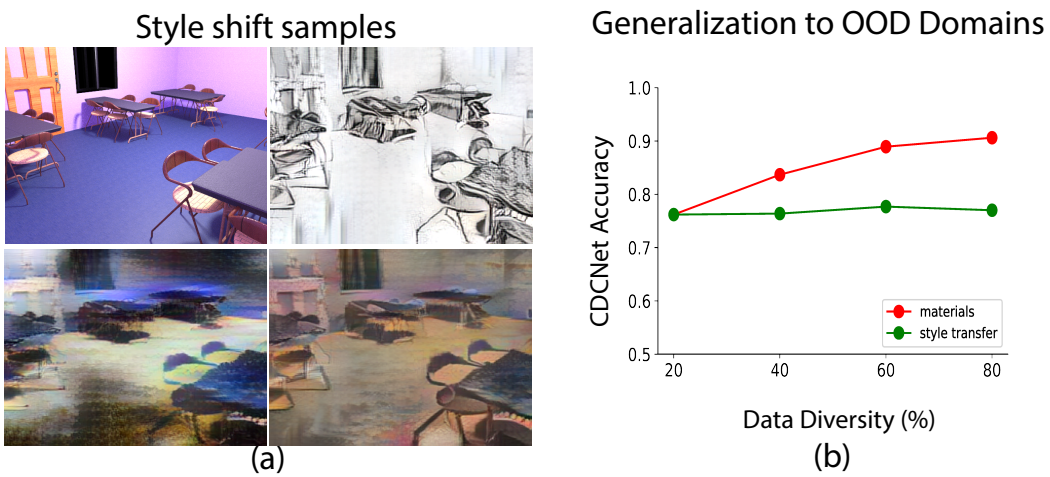



Figure S5: **Material shift vs style transfer shift**(a) Sample images from style transfer domains created by applying style transfer from AdaIn [25] to one material domain from *SemanticDG*. (b) Increasing diversity using additional materials helps the model generalize better to OOD Materials. However, there is no increase in performance as additional style transfer domains are used during training.

**Full model:** This is the full CDCNet model reported in the main paper.

As reported in Fig. S5, random chance for performance on OOD materials is 0.08 which has been tested by training images using the empty baseline protocol described above. In comparison, CDCNet trained with only context information achieves an accuracy of 0.33, which is a positive control suggesting the utility of contextual information in generalizing to the OOD Materials domain. Similarly, training a model with no context information results in an accuracy of 0.5. This dip in performance from the full model shows that removing context information hurts performance.

Another key component of our model is the contrastive loss. As shown in Fig. S5, removing the contrastive loss brings the accuracy of the model down from 0.94 to 0.89. Thus, the contrastive loss helps improve generalization significantly. Furthermore, we also find that applying contrastive loss to the features $F_{m,t,c}$ is sufficient, and that adding another loss term to both streams does not help.

## Section S5 Style transfer as an alternative to material diversity

Several existing works rely on increasing domain diversity using AdaIn [25]-based methods. These style transfer methods change the colors in the image while retaining object boundaries, but do not modify materials explicitly. Here we evaluated how well models perform if diversity is increased using style transfer as opposed to material diversity. We started with one material domain, and created four additional domains using style transfer. Thus, the total number of domains (and images) is the same as the material domains in *SematicDG*. The only difference is that instead of four additional material domains, we have four additional style transfer domains. These models are then tested on the held out OOD Materials domain. As can be seen from Fig. S5, style transfer domains do not enable models to generalize to new materials as well as material shift domains presented in *SemanticDG*. This suggests that, in order to build models that can generalize to unseen materials, we need to increase diversity using additional materials, and that style transfer is not a viable solution for this problem.

## Section S6 Hyperparameters used for all models

**CDCNet:** As our model builds on top of CRTNet [20] as backbone, we use the same hyperparameters for the backbone as reported in the original paper. All models were trained for 20 epochs with a learning rate of 0.0001, with a batch size of 15 on a Tesla V100 16Gb GPU.

**Domain generalization and Adaptation:** We used the code from Gulrajani et al. [75] to train and test domain generalization and domain adaptation methods on our dataset. The code is available here: `https://github.com/facebookresearch/DomainBed`. To begin, we ran all available models and tried 10 random hyperparameter initializations. Of these, we picked the best performing hyperparameter seed—24596. We also picked the 10 top performing algorithms as the baselines reported in the paper.

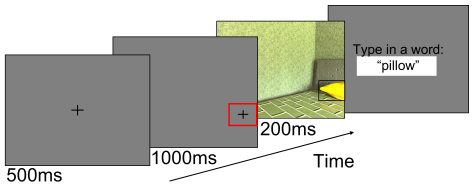



Figure S6: Subjects were presented with a fixation cross (500 ms), followed by a bounding box indicating the target object location (1000 ms). The image was shown for 200 ms. After image offset, subjects typed one word to identify the target object.

**FasterRCNN:** We used the code from Bomatter et al. [20] to train and test the modified FasterRCNN model for recognition. The code is available here: `https://github.com/kreimanlab/WhenPigsFlyContext`, and we used the exact hyperparameters mentioned in the repo.

## Section S7 Human Experiment Details

We show a schematic diagram of the human psychophysics experiments in Figure S6.

Rather than N-way categorization (e.g., [76]), we used a more unbiased probing mechanism whereby subjects could use any single word to describe the target object. We independently collected ground truth answers for each object in a separate MTurk experiment with infinite viewing time and real-world images from the ScanNet dataset. These Mturk subjects did not participate in the main experiments. Answers in the main experiments were then deemed correct if they matched any of the ground truth responses [24].

In ScanNet dataset [14], we realized that some class labels are synonyms. Thus, we also incorporated the images with synonym class labels into the stimulus set for human experiments. For example, we merged a set of images labelled as “couch" with images labelled as “sofa" and re-label both sets of images as “sofa", which is one of the ground truth class labels in the *SemanticDG* benchmark.

## Section S8 Human Experiment on the geometry domain shift

In the main paper, we covered human experiments on material, light, viewpoint domain shifts and real-world images (Table 3). Here, we added one more human experiment on geometry domain shift. The experiment schematic and procedures are the same as introduced in Section 5.2 and Section S7. The human accuracy in geometry experiment was 66.7%. Compared with CDCNet, human performance still serves as the upper bound. Surprisingly, despite not being exposed to any images from geometry domains, we found that CDCNet can generalize reasonably well in geometry domain shifts after being trained on five light domains (61%), which is very close to human performance (only 6% lower than humans).

## Section S9 Comparing generalization in humans and computer vision models

Despite all these qualitative comparison between humans and computational models, we acknowledge that a fair and direct quantitative comparison between humans and computational models in SemanticDG is difficult because of the following reasons: (i) The source domain for humans is the real world and semantic shifted domains are target domains in human experiments. The reverse holds for computational models here, where they are trained on shifted semantic domains and tested on real-world images. (ii) Past experiences differ between humans and computational models, which might influence generalization ability. Going beyond SemanticDG and ScanNet dataset, humans have accumulated decades of other experiences from the real world, while this is not the case for computational models. Indeed, we show that context could be one of these missing aspects in existing models that is critically important for generalization ability in computational models. (iii) Compared with other real-world image datasets, e.g., [29] where humans almost achieve perfect recognition performance in its generalization tests, the absolute human performance is around 66% in SemanticDG, which is far from perfect. It is possible that humans are not biased by the set of selected object classes from SemanticDG, while the computational models learn these inductive biases from the training set. (iv) The absolute human performance on the three semantic domains (lights, materials and viewpoints) is not as impressive as the performance of computational models from Figure 3. This could be due to the way that we structured the training and test set for the computational models. In the experiments in Sec 6.1, train and test images are constructed with one-to-one correspondence in a controllable manner. That is, except for the semantic domain we are varying (e.g., material),

the other aspects of the semantic scene (e.g., room layouts) remain exactly the same between train and test sets. However, in human behavioral experiments, we do not include these training sets.

Despite all these caveats, it is instructive to show results for humans and models on the same images in the SemanticDG benchmark on Table 3. We tried to mitigate the differences in training by focusing on the qualitative impact of generalization across various domains compared to the real-world condition.

## Section S10 Source Code and Dataset Link (anonymous)

Anonymous link for source code and *SemanticDG* benchmark: `https://drive.google.com/drive/folders/1NBiiUPtgCUA-FoolOZ9id_Xor7eyzrUL?usp=sharing`