Improving out-of-distribution generalization by mimicking the human visual diet

Anonymous Author(s)
Affiliation
Address
email

Abstract

Human visual experience is markedly different from the large scale computer vision datasets constructed by scraping the internet. Babies densely sample a few $3D$ scenes with diverse variations, while datasets like ImageNet contain one single snapshot from millions of 3D scenes. We investigated how these differences in input data composition (*ie.,* visual diet) impact the Out-Of-Distribution (OOD) generalization capabilities of a visual system. We found that training models on a dataset mimicking attributes of the human-like visual diet improved generalization to OOD lighting, material, and viewpoint changes by up to 18%. This was true despite being trained on $1,000$-fold lesser training data. Furthermore, when trained on purely synthetic data and tested on natural images, incorporating these attributes in the training dataset improved OOD generalization by 17%. These experiments are enabled by our newly proposed benchmark—the Human Visual Diet (HVD) dataset, and a new model (Human Diet Network) designed to leverage the attributes of a human-like diet. These findings highlight a critical problem in modern day Artificial Intelligence—building better datasets requires thinking beyond dataset size, and improving data composition. All data and source code are available at `https://bit.ly/3yX3PAM`.

1 Introduction

The development of the human visual system is intricately tied to the visual experiences encountered from infancy [1, 2, 3, 4, 5, 6, 7, 8, 9, 6, 7]. These visual experiences are constrained by the structure of the spaces we occupy, resulting in data significantly different from large-scale datasets used in computer vision. **Fig. 1(a)** illustrates two such differences. First, children learn from the physical space they occupy—a few 3D scenes and objects viewed under diverse real-world transformations including viewpoints, lighting, object textures, and natural occlusions. Second, children always view objects in the context of their surroundings. We refer to these as *real-world transformational diversity (RWTD)* and *scene context*, respectively. Here, we investigate how these differences in input data composition impact Out-Of-Distribution (OOD) generalization performance.

We found that incorporating these attributes into the training data significantly improves generalization. Models trained with a human-like visual diet achieve up to 18% improved performance on OOD lighting, materials, and viewpoint changes. In fact, training with such data outperforms training models on 1000-fold larger internet-scraped datasets. These experiments are enabled by two key technical contributions. First, the **Human Visual Diet (HVD)** dataset, which contains both transformational diversity and scene context [10, 11] (**Figure Sup1**). Second, the **Human Diet Network (HDNet)** model designed to leverage the attributes present in HVD (See **Fig. 1(c)**). HDNet exploits transformational diversity by employing a contrastive loss over real-world transformations (lighting, material, $3D$ viewpoint changes), and uses a two-stream architecture to jointly reason over target

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

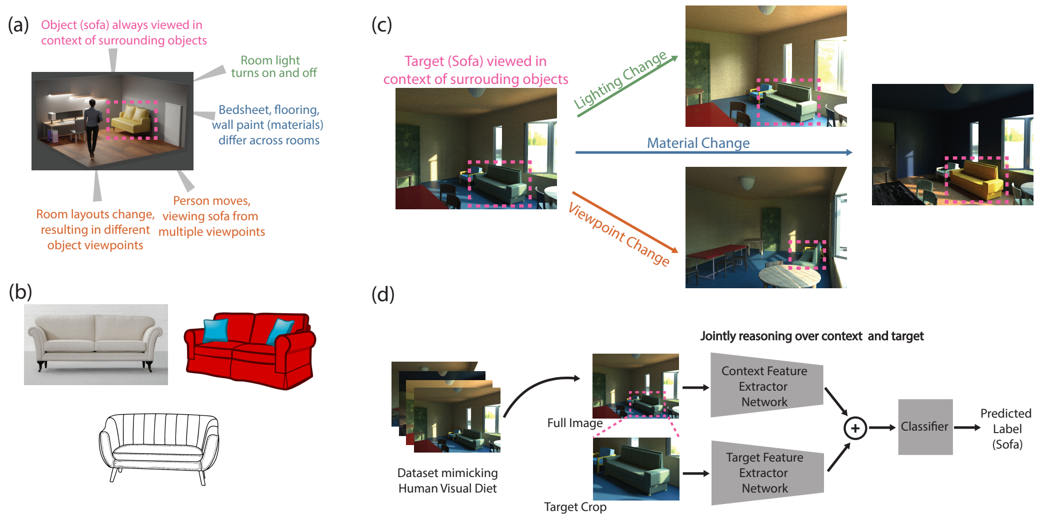


Figure 1: **Mimicking the human visual diet.** (a),(b) Comparing human and machine visual diets: The desk in the 3D room is viewed under a variety of real-world transformations, and objects are seen in the context of their surroundings. Both attributes are missing in internet scraped images of desks. (c) Human Visual Diet (HVD) dataset contains images with disentangled lighting, material, and viewpoint changes to a 3D scene where objects are shown in context. (d) Human Diet Network (HDNet) leverages these attributes by using a two-stream architecture which reasons over both target object and its surrounding scene context, and uses a contrastive loss over real-world transformations.

and scene context to perform context aware visual recognition. We add to a growing body of works positing the importance of mimicking the human visual diet [11, 7, 6, 12, 10, 13, 14] by extending them, and showcasing the improved OOD generalization resulting from such training data.

2 Related Work

Out-of-Distribution (OOD) generalization continues to be the Achilles heel of Modern AI [15, 16, 17]. Failure modes include OOD rotations and translations [15, 16, 17], real-world transformations including 3D viewpoints [18, 19, 20, 21, 22, 23], changes in lighting [21, 24, 25], and color changes [26, 27], among other transformations. Existing approaches to counter this include—specialized architectures [28, 29, 30, 31, 32, 33, 34], novel pre-processing and data augmentation strategies [35, 36, 37, 38, 39], and generative modeling [40, 41], among others. Lately, practitioners have made datasets larger than ever in the hopes that billion scale datasets like LAION-5B [42] and IG-1B Targeted [43] will contain enough information to leave very little out of the distribution. However, despite unprecedented progress, OOD samples remain an unsolved problem [44, 45, 46]. In contrast, some recent work has emphasized the importance of training with more human like data [8, 9, 6, 7]. This includes incorporating scene context [47], temporal structure [12], binocular vision [48, 49], and goal-directed/active sampling [14, 13, 50, 51, 52], among others. Our work extends these to Out-of-Distribution generalization.

3 Datasets with controlled variations in lighting, materials and viewpoints

We present three new benchmarks for measuring OOD generalization across real-world transformations in lighting, materials, and viewpoint changes.

3.1 Human visual diet (HVD) Dataset

1,288 $3D$ scenes from ScanNet [53] were reconstructed using the OpenRooms framework [54, 55], and 15 photo-realistic domains were constructed with these scenes by introducing 3 real-world transformations—lighting, material, and viewpoint changes. For each domain, $19,800$ images were rendered resulting in a total of $300,000$ images containing 1 million object instances with controlled

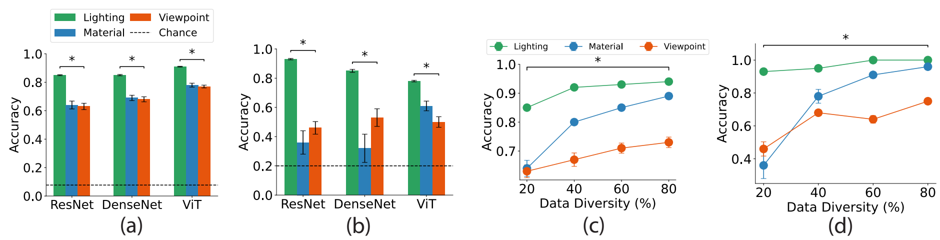


Figure 2: **Real-world transformational diversity significantly improved generalization.** (a) Models struggle to generalize across real-world transformations—especially material and viewpoint changes for HVD, and (b) for Semantic-iLab. (c) Generalization improves significantly as real-world transformational diversity (RWTD) is increased for HVD, and (d) for Semantic-ilab.

variations in lighting, object materials, and viewpoints (see **Fig. Sup1(a)**). Additional details on the construction of OOD material, viewpoint and lighting domains are provided in Sec. Sup1.

3.2 Semantic-iLab dataset

Images from iLab [56] were modified to create a natural image dataset with variations in lighting, material and viewpoints (**Fig. Sup1 (b)**). iLab contains objects from 15 categories placed on a turntable and photographed from varied viewpoints. Fist, a foreground detector was used to extract the object. Then, material variations were implemented using AdaIN [57] based style transfer on these object masks and the style transferred object was overlayed onto the original background. Lighting changes were simulated by modifying the white balance. Unlike HVD, this dataset does not contain scene context. Additional details can be found in supplementary **Sec. B**.

3.3 Syn2Real dataset: Natural image test set from ScanNet

The Syn2Real dataset is composed of a test set of natural images from the ScanNet dataset, and a training set of only synthetic images from HVD. The natural image test set was created by annotating images from ScanNet [53]. To capture distinct images, one frame was sampled every 100 frames from ScanNet's raw video footage. These frames were then annotated using LabelMe.

4 Human Diet Network (HDNet)

A schematic of the proposed HDNet is shown in **Fig Sup5**. Given the training dataset $D = \{x_i, y_i\}_{i=1}^{n}$, HDNet is presented with an image x_i with multiple objects and the bounding box for a single target object location. The target $(I_{i,t})$ 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 $I_{i,t}$. Inspired by the eccentricity dependence of human vision, HDNet 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. We also utilize contrastive learning over real-world transformations—Samples of the same object category (but different lighting, 3D viewpoint, or texture) serve as positive pairs, while samples of different object category serve as negative pairs. Additional details on the model are provided in Sec. D.

5 Results

One domain per transformation was held out as the OOD test set and never used for training. As Real-World Transformational Diversity (RWTD) was increased from 1 to 4 domains (corresponding to 20% to 80% data diversity), the number of images sampled per domain were reduced. This ensured a fixed training dataset size. All models were pre-trained on ImageNet.

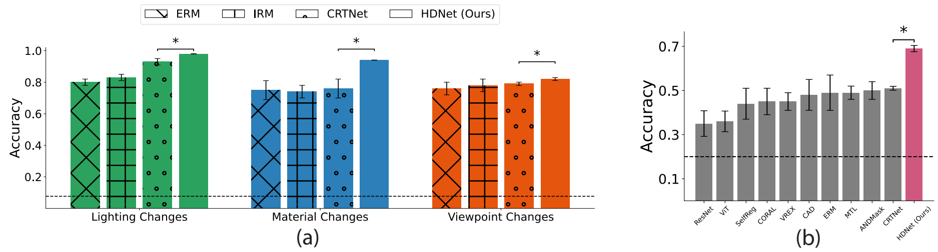


Figure 3: **Scene Context improves OOD generalization.** (a) HDNet explicitly leverages scene context resulting in substantially better generalization than domain generalization approaches like ERM [61] and IRM [30] for all three transformations (lighting, material, and viewpoint changes). (b) Human-like visual diet enables improved generalization from synthetic to natural image data.

Real-World Transformation	AND Mask [28]	CAD [34]	COR AL [29]	MTL [61]	Self Reg [31]	VREx [33]	Faster RCNN [62]	HDNet (ours)
Light	0.82	0.80	0.81	0.81	0.75	0.83	0.95	**0.98**
Materials	0.75	0.75	0.75	0.74	0.74	0.75	0.78	**0.94**
Viewpoints	0.75	0.77	0.79	0.79	0.76	0.78	0.65	**0.83**

Table 1: **Contextual information improves OOD generalization.** All models were trained with 80% transformational diversity and tested on the held-out 20%. HDNet beats all specialized domain generalization baselines and a FasterRCNN modified to do object recognition, by a large margin.

5.1 Models with low diversity and minimal context struggle to generalize.

Fig.2 presents generalization performance of models trained with low transformational diversity and minimal scene context—data was sampled from only 1 domain, and images were cropped to show only the target object. This diet is representative of internet scraped datasets like ImageNet [58], and these models served as a lower baseline to quantify the impact of a human-like visual diet.

For HVD (**Fig. 2(a)**), ResNet18 generalized better across lighting changes than material changes (two-sided t-test, $p < 10^{-5}$) or viewpoint changes (two-sided t-test, $p < 10^{-6}$). There is ample room for improvement, especially when tested on OOD material and viewpoints. Similar conclusions can be drawn for DenseNet [59] and ViT [60] architectures. For Semantic-iLab (**Fig. 2(b)**) as well, ResNet18 generalized better across OOD lighting than OOD materials (two-sided t-test, $p < 10^{-6}$) or OOD viewpoints (two-sided t-test, $p < 10^{-6}$). In the Semantic-iLab dataset, the degree of generalization for material and viewpoints were particularly low. These conclusions held true for DenseNet and ViT as well. In sum, models trained with minimal diversity and context showed only moderate generalization, especially struggling with material and viewpoint changes.

5.2 Utilizing real-word transformational diversity (RWTD) improves generalization

OOD Generalization improved with transformational diversity for all three transformations in the HVD dataset (**Fig. 2(c)**). For lighting: 0.85 to 0.94, $p < 10^{-6}$; material: 0.64 to 0.89, $p < 10^{-5}$; viewpoint: 0.63 to 0.73, $p < 10^{-6}$. This improvement was significantly greater for OOD materials than for OOD lighting ($p < 10^{-4}$) and OOD viewpoints ($p < 10^{-4}$). Transformational diversity improved generalization for the Semantic-iLab dataset as well (**Fig. 2(d)**). For lighting: 0.93 to 1.0, $p < 10^{-3}$; materials: 0.36 to 0.96, $p < 10^{-4}$; viewpoint: 0.46 to 0.75, $p < 10^{-7}$. As with the HVD dataset, improvement in generalization was higher for unseen materials than for unseen lighting ($p < 10^{-3}$) and unseen viewpoints ($p < 10^{-6}$). Thus, OOD generalization improved across all real-world transformations with transformational diversity. Inn fact, with sufficient diversity, generalization to OOD lighting and materials reached almost ceiling levels. However, despite improvement, OOD viewpoints remained a challenge.

Real World Transformation	Dino V2	ResNet50 SWSL	ResNet18 SWSL	ResNext101 32x4d SWSL	ResNext101 32x16d SWSL	ResNext50 32x4d SWSL	HDNet (Ours)
Light	0.94	0.9	0.88	0.93	0.93	0.91	**0.98**
Materials	0.79	0.73	0.67	0.77	0.79	0.74	**0.94**
Viewpoints	0.74	0.72	0.65	0.74	0.78	0.73	**0.83**

Table 2: **Our approach beats models trained with 1000x more data.** HDNet was pre-trained on ImageNet and finetuned on data with both transformational diversity and scene context. Baselines were pre-trained on 1000-fold more data, but fine-tuned on data not containing these two attributes. HDNet beats all baselines by a large margin for all three transformations, despite being trained on 1000-fold smaller training data.

5.3 Utilizing scene context improves generalization.

We compared HDNet with a suite of baselines that do not utilize scene context. This includes domain generalization (DG) architectures, and a modified FasterRCNN model designed to perform visual recognition. We also added a recent context-aware model (CRTNet [63]) to the comparison. All models were trained with 80% Transformational Diversity, i.e., 4 training domains. HDNet beat all DG methods with statistical significance (two-sided t-test, $p < 0.05$) for all three transformations. Top three baselines are presented in **Fig. 2**(e). The remaining baselines are shown in **Table 1**. The best performing baseline was another context-aware model—CRTNET [63]. HDNet outperformed all benchmarks on all three transformations. In summary, approaches utilizing scene context (HDNet and CRTNet) outperformed all specialized DG approaches on all real-world transformations, and our proposed HDNet also outperformed the closest baseline (CRTNet). We present several additional experiments on the role of scene context in the supplement in Sec. F.

5.4 Human-like visual diet outperforms billion-scale internet-scraped datasets

Next, we compared HDNet with visual recognition models trained with 1,000x more data (**Table. 2**). All models except HDNet were pre-trained on the IG-1B dataset [43], and then fine-tuned on data with 20% RWTD and with object crops *ie.,* low transformational diversity and minimal context. In comparison, HDNet was pre-trained on ImageNet and fine-tuned with data consisting of 80% RWTD and scene context *ie.,* human-like visual diet. All models were fine-tuned on the same number of images. HDNet outperformed all billion-scale baselines by large margins despite being trained on 1000x less data (**Table. 2**, two-sided t-test, $p < 0.001$).

5.5 Human-like visual diet enables generalization to real-world images

HDNet trained with RWTD and scene context achieved an accuracy of 0.69, while the best baseline (IRM [30]) trained without a human-like diet achieved an accuracy of 0.51 (**Fig. 3**(**b**)). Thus, incorporating these attributes into the training dataset enabled HDNet to generalize significantly well from a purely synthetic training data to a natural image test set (two-sided t-test, $p < 0.05$).

6 Conclusions

We investigated the impact of data composition on the out-of-distribution generalization capabilities of visual recognition models. Specifically, we demonstrated that incorporating two key components of the human visual diet—transformational diversity and scene context improve generalization to OOD viewpoints, lighting, and material changes. Our contributions include three new benchmarks, and a novel architecture that model and leverage these human-like visual attributes. This work provides an approach complementary to existing directions on data augmentation and specialized domain generalization architectures. While our results are promising, the human visual diet is complex and multifaceted, with several additional features like temporal information, egocentric views, embodiment, and goal-driven/active sampling warranting future investigation. We believe this work opens new avenues for aligning biological and artificial vision systems, and advancing generalization in Artificial Intelligence.

NeurIPS paper checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction state the main claims, approach and the experiments support the claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: They are provided in the conclusions section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We have no Proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details are provided alongside code and data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and code are provided and are free for anyone to use.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, all information are provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Yes, we used two-sided t-tests for statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, details are provided in experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and reviewed the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no societal imapct of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work raises no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: No such assets were used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Dataset comes with details on how to use it.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

References

[1] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[2] Gabriel Kreiman. *Biological and Computer Vision*. Cambridge University Press, 2021.

[3] Michael J Arcaro, Peter F Schade, Justin L Vincent, Carlos R Ponce, and Margaret S Livingstone. Seeing faces is necessary for face-domain formation. *Nature neuroscience*, 20(10):1404–1412, 2017.

[4] David H Hubel and Torsten N Wiesel. Effects of monocular deprivation in kittens. *Naunyn-Schmiedebergs Archiv für Experimentelle Pathologie und Pharmakologie*, 248:492–497, 1964.

[5] NW Daw and HJ Wyatt. Kittens reared in a unidirectional environment: evidence for a critical period. *The Journal of physiology*, 257(1):155–170, 1976.

[6] Justin N Wood and Samantha MW Wood. The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, 42(4):1391–1406, 2018.

[7] Justin N Wood and Samantha Marie Waters Wood. The development of object recognition requires experience with the surface features of objects. *bioRxiv*, pages 2022–12, 2022.

[8] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. *Advances in neural information processing systems*, 31, 2018.

[9] Donsuk Lee, Pranav Gujarathi, and Justin N Wood. Controlled-rearing studies of newborn chicks and deep neural networks. *arXiv preprint arXiv:2112.06106*, 2021.

[10] Saber Sheybani, Zoran Tiganj, Justin N. Wood, and Linda B. Smith. Slow change: An analysis of infant egocentric visual experience. *Journal of Vision*, 23(9):4685–4685, Aug 2023.

[11] Linda B Smith and Lauren K Slone. A developmental approach to machine learning? *Frontiers in psychology*, 8:296143, 2017.

[12] Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] John Tsotsos, Iuliia Kotseruba, Alexander Andreopoulos, and Yulong Wu. Why does data-driven beat theory-driven computer vision? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[14] John K Tsotsos. On the relative complexity of active vs. passive visual search. *International journal of computer vision*, 7(2):127–141, 1992.

[15] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. 2018.

[16] 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.

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

[18] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[19] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. *arXiv preprint arXiv:1808.02651*, 2018.

[20] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4302–4311, 2019.

[21] 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.

[22] 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 invariant neurons and generalization to out-of-distribution orientations and illuminations. *Neural Networks*, 155:119–143, 2022.

[23] Kaiyu Zheng, Anirudha Paul, and Stefanie Tellex. A system for generalized 3d multi-object search. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1638–1644, 2023.

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

[25] Qian Zhang, Qing Guo, Ruijun Gao, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Adversarial relighting against face recognition. *arXiv preprint arXiv:2108.07920*, 2021.

[26] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4773–4783, 2019.

[27] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1160, 2020.

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

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

[35] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[36] 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.

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

[38] Spandan Madan, Zoya Bylinskii, Carolina Nobre, Matthew Tancik, Adria Recasens, Kimberli Zhong, Sami Alsheikh, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Parsing and summarizing infographics with synthetically trained icon detection. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pages 31–40, 2021.

[39] Spandan Madan, Zoya Bylinskii, Carolina Nobre, Matthew Tancik, Adria Recasens, Kimberli Zhong, Sami Alsheikh, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Parsing and summarizing infographics with synthetically trained icon detection. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE, April 2021.

[40] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020.

[41] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6678–6687, 2020.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[43] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[45] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

[46] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning, 2021. *URL https://arxiv. org/abs/2111.10050*.

[47] 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.

[48] Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33:9960–9971, 2020.

[49] A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3):271–283, 2024.

[50] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42:177–196, 2018.

[51] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

[52] Madeline Helmer Pelgrim, Ivy Xiao He, Kyle Lee, Falak Pabari, Stefanie Tellex, Thao Nguyen, and Daphna Buchsbaum. Find it like a dog: Using gesture to improve object search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[53] 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.

[54] 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.

[55] 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.

[56] Ali Borji, Saeed Izadi, and Laurent Itti. ilab-20m: A large-scale controlled object dataset to investigate deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2221–2230, 2016.

[57] 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.

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

[59] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[60] 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.

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

[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] 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.

[64] 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.

[65] 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.

[66] Kentaro Wada. labelme: Image polygonal annotation with python. `https://github.com/wkentaro/labelme`, 2018.

[67] 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.

[68] 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.

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

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

[71] 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.

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

[73] 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.

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

[75] 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.

[76] 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.

[77] 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.

[78] Joost CF De Winter. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18(1):10, 2019.

[79] Harry O Posten. Two-sample wilcoxon power over the pearson system and comparison with the t-test. *Journal of Statistical Computation and Simulation*, 16(1):1–18, 1982.

Supplementary Materials

A Details on the construction of HVD domains

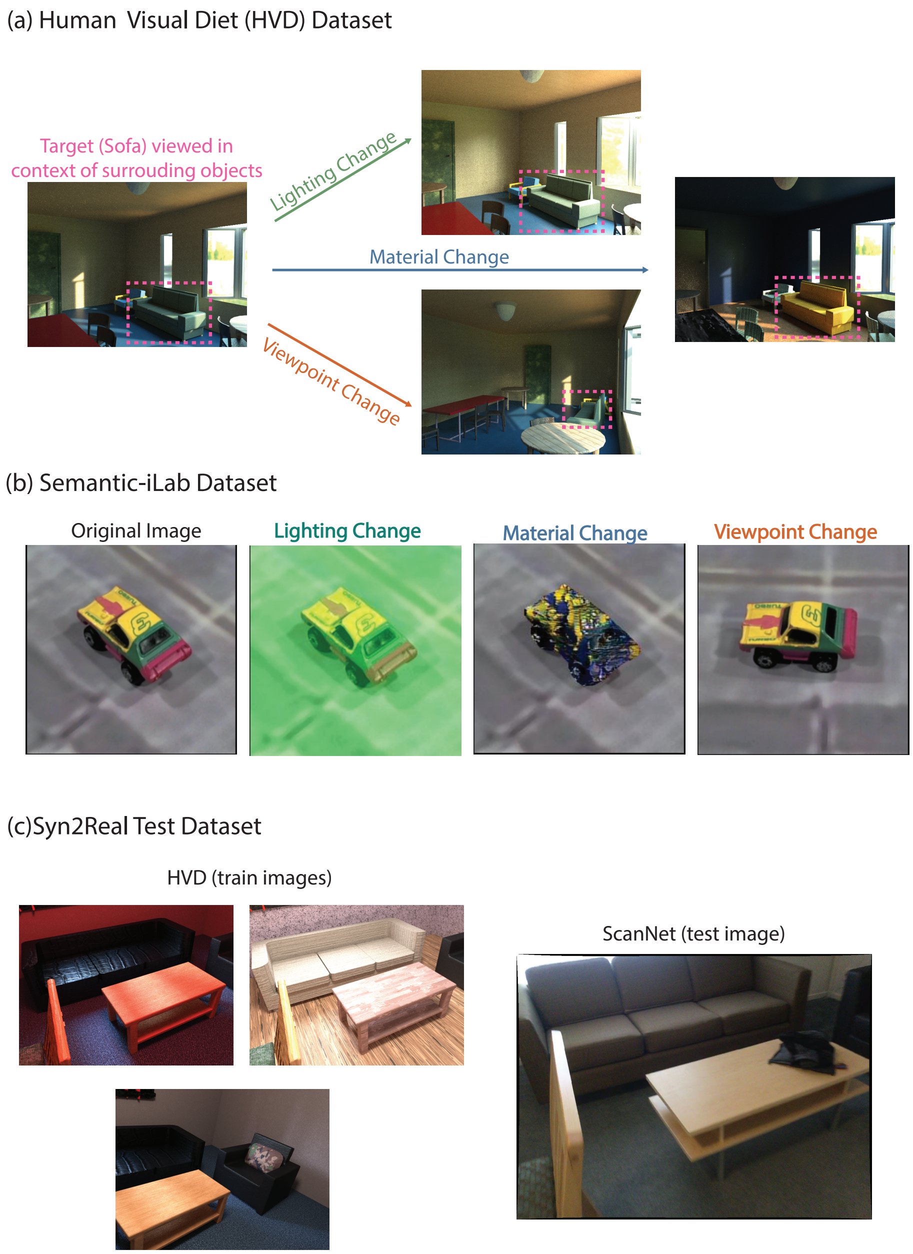


Figure Sup1: **Datasets with real-world transformations.** (a) Sample images from the Human visual diet dataset: We created 15 photo-realistic domains with three, disentangled real-world transformations—lighting, material, and viewpoint changes. Each 3D scene was created by reconstructing an existing ScanNet [53] scene using the OpenRooms framework [54], followed by introduction of controlled changes in scene parameters before rendering these images. (b) Sample images from the Semantic-iLab dataset: We modify the existing iLab dataset [56] by augmenting images with changes in lighting and material. These changes are achieved by modifying the white balance and using AdaIN [64] based style transfer, respectively. (c) Syn2Real dataset constructed with paired 3D scenes—synthetic images for training and natural images for testing.

Figure Sup2: *Example images showing lighting tranformations.* We show paired images from different lighting transformation domains between the right and left column in each row. All other parameters held constant.

Figure Sup3: *Example images showing material tranformations.* We show paired images from different material transformation domains between the right and left column in each row. All other parameters held constant

Figure Sup4: *Example images showing viewpoint tranformations.* We show paired images from different viewpoint transformation domains between the right and left column in each row. All other parameters held constant

A.1 Lighting, Material, and Viewpoint domains:

Material shift domains: We used 250 high quality, procedural materials from Adobe Substances including different types of wood, fabrics, floor and wall tiles, and metals, among others. These were split into sets of 50 materials each to create 5 different material domains (supplementary **Fig. Sup3**). For each domain, its 50 materials were randomly assigned to scene objects. One domain was held out for testing (OOD Materials), and never used for training any model.

Light shift domains: Outdoor lighting was controlled using 250 High Dynamic Range (HDR) environment maps from the Laval Outdoor HDR Dataset [65] and OpenRooms, which were split into 5 sets of 50 each (one set per domain). Disjoint sets of indoor lighting were created by splitting the HSV color space into chunks of disjoint hue values. Each domain sampled indoor light color and intensity from one chunk (supplementary **Fig. Sup2**). One domain was held out for testing (OOD Light), and never used for training.

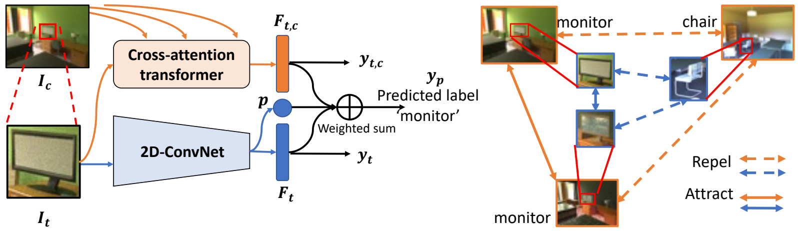
Viewpoint shift domains: Controlling object viewpoints presents a challenge as indoor objects are seen across a variety of azimuth angles (i.e., side vs front) across 3D scenes. Thus, to create disjoint viewpoint domains (supplementary **Fig. Sup4**) we chose to control the zenith angle by changing the height at which the camera is focusing. Again, of the 5 domains, one was held out for testing (OOD Viewpoints). We show sample images from the *Semantic iLab* dataset in Fig. Sup1(b) created by modifying the existing iLab [56] dataset. This is a multi-view dataset, and hence already contains viewpoint shifted variations of the same objects. We modify the dataset to also contain material and light shifts. To mimick light shift, we modified the white balance of the original images, as shown in Fig. Sup1(b)(b). For material shifts, we first run a foreground detector on these objects using Google's Cloud Vision API. We also run style transfer on these images using AdaIn [57]. Then, we overlay the style transferred image on to the object mask on the original image to mimick material shifts. Note that this is approximate, and does not model the physics of material transfer in the same way as our rendered HVD dataset which is far more photorealistic, as shown in Fig. Sup3. Material shifted *Semantic iLab* images are shown in Fig. Sup1(b)(c). As the dataset is originally multi-view, we do not need to generate new viewpoints and can use images of a different viewpoint from the original dataset as shown in Fig. Sup1(b)(d).

A.2 Sample images from the HVD Dataset

We present additional images from the HVD dataset. Each figure shows change in one scene parameter, while holding all others constant. In Fig. Sup2 we show images from two different light domains. Note that the first three rows in Fig Sup2 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. Sup3 shows five different scenes from two training domains with a material shift. Fig. Sup4 shows viewpoint shifted domains.

B Details on the construction of the Semantic iLab dataset

We show sample images from the *Semantic iLab* dataset in Fig. Sup1(b) created by modifying the existing iLab [56] dataset. This is a multi-view dataset, and hence already contains viewpoint shifted variations of the same objects. We modify the dataset to also contain material and light shifts. To mimick light shift, we modified the white balance of the original images, as shown in Fig. Sup1(b). For material shifts, we first run a foreground detector on these objects using Google's Cloud Vision API. We also run style transfer on these images using AdaIn [57]. Then, we overlay the style transferred image on to the object mask on the original image to mimick material shifts. Note that this is approximate, and does not model the physics of material transfer in the same way as our rendered HVD dataset which is far more photorealistic, as shown in Fig. Sup3. Material shifted *Semantic iLab* images are shown in Fig. Sup1(b). As the dataset is originally multi-view, we do not need to generate new viewpoints and can use images of a different viewpoint from the original dataset as shown in Fig. Sup1(b).



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

Figure Sup5: **Architecture overview for the Human Diet Network(HDNet)**. (a) Modular steps carried out by HDNet in context-aware object recognition. HDNet consists of 3 modules: feature extraction, integration of context and target information, and confidence-modulated classification. HDNet 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. HDNet 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 HDNet 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 (**Sec. 3.1**).

C Details on the construction of the Syn2Real dataset

We made three adaptations for these experiments. Firstly, as both ScanNet and ImageNet contain natural images and overlapping categories, we trained models from scratch to ensure pre-training does not interfere with our results. Thus, these models never saw any real-world images, not even ImageNet as they were not pretrained on those datasets. Secondly, we trained and tested models on overlapping classes between HVD and ScanNet. Finally, we used the LabelMe [66] software to manually annotate a test set from ScanNet and training set for the HVD dataset using the same procedure to make sure biases from the annotation procedure do not impact experiments. Thus, all models were trained purely on synthetic data from HVD and tested on only real-world natural image data from ScanNet as shown in **Fig. Sup1(c)**.

D Details on the Human Diet Network

The context stream is a transformer decoder, and the network integrates object and context information via hierarchical reasoning through a stack of cross-attention layers in the transformer. This allows HDNet to be more robust under distribution shifts in object context. Furthermore, HDNet utilizes a contrastive learning method on 3D transformations.

A model that always relies on context can make mistakes under distribution shifts. Thus, to increase robustness, HDNet 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, HDNet 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.

Contrastive learning has benefited many applications in computer vision tasks (*e.g.*, [67, 68, 69, 70, 31]). 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 contrastive learning method on 3D transformations.

Our contrastive learning framework builds on top of the supervised contrastive learning loss [71]. 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 $\tilde{y}_{2k} = \tilde{y}_{2k-1} = \tilde{y}_k$. Domain shifts are randomly selected from a set of HVD 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 is:

$$L_{contrast} = \sum_{m \in M} L_m = \sum_{m \in M} \frac{-1}{|P(m)|} \sum_{p \in P(m)} \log \frac{\exp(z_m \cdot z_p / \tau)}{\sum_{a \in A(m)} \exp(z_m \cdot z_a / \tau)} \quad \text{(Sup1)}$$

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 HDNet 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 [63, 47], contrastive learning on the context-modulated object representations enforces HDNet to learn generic category-specific semantic representations across various domains. τ is a scalar temperature value which we empirically set to 0.1.

Overall, HDNet 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 $I_{m,t}$, HDNet 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:

$$L = \alpha L_{contrast,c,t} + L_{classi,t} + L_{classi,p} + L_{classi,c,t} \quad \text{(Sup2)}$$

Hyperparameter α is set to 0.5 to balance the supervision from constrastive learning and the classification loss. Supplementary Table Sup2 shows that the contrastive loss introduced in HDNet results in improved performance across all real-world transformations.

E Additional experiments with real-world transformational diversity

E.1 Real-world transformations outperform traditional data augmentation.

We investigated how real-world transformational diversity (RWTD) compares to traditional data augmentation strategies including 2D rotations, scaling, and changes in contrast. Models trained with a visual diet consisting of 80% RWTD were reported in **Fig.3(e)**. We compared these with models trained with a visual diet consisting of 20% RWTD + traditional augmentation. As before, all models were tested on unseen lighting, material, and viewpoint changes.

The number of training images was kept constant across all training scenarios to evaluate the quality of the training images rather than their quantity. Training set size equalization was achieved by sampling fewer images per domain in the 80% RTWD training set. For instance, for HVD experiments with unseen viewpoints we sampled $15,000$ training images per viewpoint domain to construct the training set with 20% RWTD + Data Augmentations. In comparison, we sampled only $3,750$ per viewpoint domain to construct the 80%RWTD training set. Thus, the initial sizes of the 80%RWTD and the 20%RWTD+Data Augmentation training sets was identical. However, due to data augmentations being stochastic the total number of unique images shown to models trained with data augmentations was much larger. Assuming a unique image was created by data augmentation in every epoch, over 50 epochs the dataset size would be 50 times larger with data augmentations. Additional details on dataset construction can be found in the methods in Methods.

HDNet trained on HVD with 80% RWTD outperformed the same architecture trained with 20% RWTD+traditional data augmentation for lighting changes (two-sided t test, $p < 10^{-4}$), material changes (two-sided t test, $p < 10^{-5}$), and viewpoint changes (two-sided t test, $p < 10^{-6}$) (**Fig. Sup6(a)**). Similar conclusions were reached for the Semantic-iLab dataset. A ResNet model trained with 80% RWTD outperformed the same architecture trained with 20% RWTD+traditional data augmentation for lighting changes (two-sided t test, $p < 10^{-4}$), material changes (two-sided t test, $p < 10^{-7}$), and viewpoint changes (two-sided t test, $p < 10^{-5}$) (**Fig. Sup6**(**b**)).

Traditional data augmentation largely involves 2D affine operations (crops, rotations) or image-processing based methods (contrast, solarize) which are not necessarily representative of real-world transformations. In summary, the positive impact of a visual diet consisting of diverse lighting, material, and viewpoint changes (real-world transformational diversity) cannot be replicated by using traditional data augmentation applied to the dataset after data collection—diversity must be ensured at the data collection level.

E.2 Real-world transformations outperform augmentation with generative AI.

Several existing works rely on increasing data diversity using AdaIn-based methods [64, 72]. These style transfer methods change the colors in the image while retaining object boundaries, but do not modify materials explicitly as done in our HVD dataset. 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. Sample images of style transfer domains are shown in **Fig. Sup6**(**c**). Corresponding images from the HVD dataset with real-world transformation in materials can be seen in **Fig. Sup1**(**a**). The total number of domains (and images) created using style transfer was kept the same as the material domains in HVD. The only difference in the training data was that instead of four additional material domains, we have four additional style transfer domains. We compared models trained with these two different visual diets—one consisting of four material domains, and the other consisting of four style transfer domains. All models were then tested on the same held-out OOD Materials domain. Style transfer domains did not enable models to generalize to new materials as well as the material shift domains presented in HVD (**Fig. Sup6**(**d**)).

These experiments support the notion that in order to build visual recognition models that can generalize to unseen materials, it is important to explicitly increase diversity using additional materials at the time of training data collection. The impact of diverse materials cannot be replicated by using style transfer to augment the dataset after data collection.

E.3 Each individual real-world transformation is helpful

Some real-world transformations are easier to capture than others. For instance, capturing light changes during data collection might be significantly easier than collecting multiple possible room layouts, or object viewpoints. Thus, it would be beneficial if training with one transformation (*e.g.,* light changes) can improve performance on a different transformation (*e.g.,* viewpoint changes). We refer to such a regime as *assymetric diversity*—as models are trained with one kind of diversity, and tested on a different kind of diversity (**Fig. Sup6**(**e**),(**f**)). In all cases, the best generalization performance was obtained when training and testing with the same real-world transformation for both HVD (**Fig. Sup6**(**e**)) and Semantic-iLab datasets (**Fig. Sup6**(**f**)). In most cases, there was a drop in performance of 10% or more when training in one transformation and testing with a different (assymetric) transformation. These experiments imply that to build models that generalize well, it is important to collect training data with multiple real-world transformations.

F Additional experiments for the role of context

Given the success of HDNet, we asked whether implementing a two-stream separation of target and context would also improve performance for other architectures. We modified ResNet18 [73] and ViT [60] to leverage scene context in the same way as HDNet. For ResNet, a two-stream version was made where each stream is a ResNet backbone. One stream operates on the target, and the other one on the scene context. Output features from each stream were concatenated, and passed through a fully connected layer for classification as shown in **Fig. 1**(**c**). The two-stream architecture for ViT was analogous. In contrast, the one-stream architecture did not use scene context and operated on the target object alone (see methods for additional details). The two-stream architectures consistently led to improved performance (two-sided t test, $p < 0.05$), as shown in **Table Sup1**.

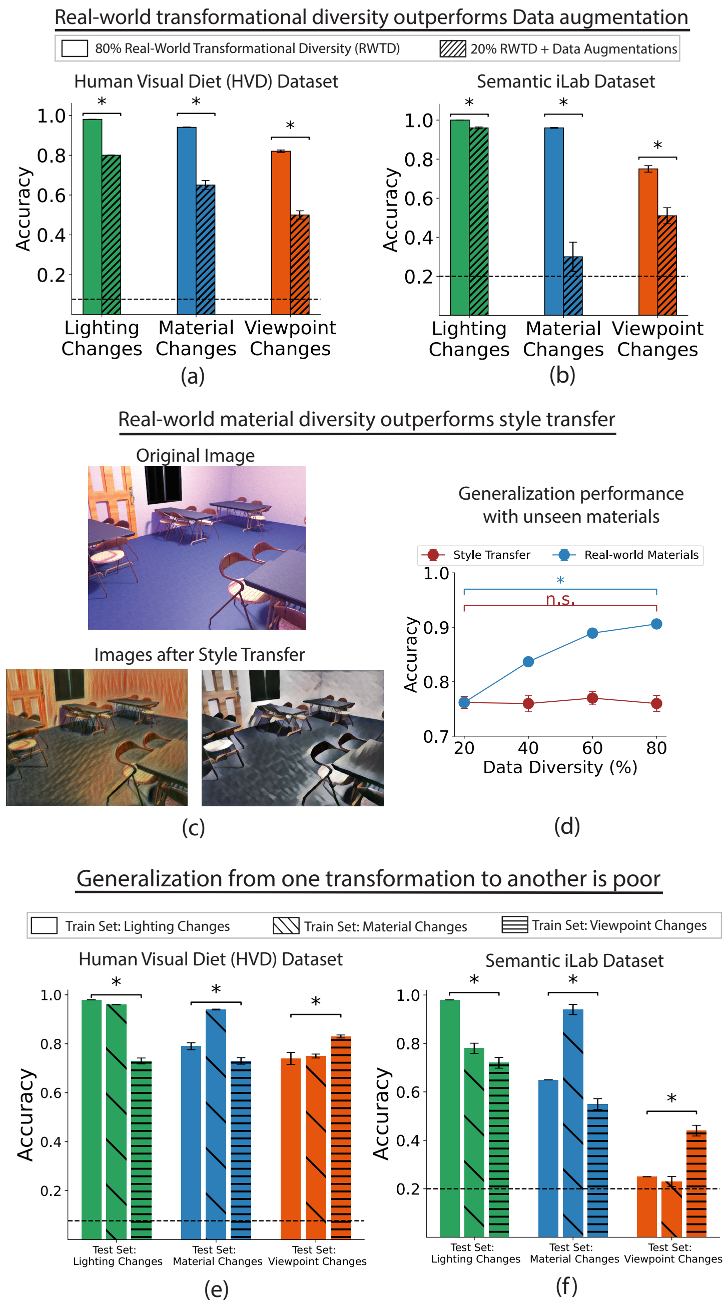


Figure Sup6: **Data post-processing does not match gains from collecting data mimicking the human visual diet**. (a),(b) Models trained with 80% real-world transformational diversity (RWTD) outperform those trained with 20% RWTD and traditional data augmentation for all transformations (lighting, material, and viewpoint) across both HVD and Semantic-iLab datasets. Number of images is held constant in these experiments. (c) Sample images from style transfer domains created using AdaIn [64]. (d) Models trained on style transfer domains generalize significantly worse than those trained with material diversity. (e),(f) Asymmetric diversity does not help generalization as much as training with the correct transformation—generalization to unseen materials is best when material diversity is added during training, as opposed to adding light or viewpoint diversity during training. Same result holds for lighting and viewpoint transformations.

Real-World Transformation	Architecture	1 Stream	2 Stream
Lighting	ResNet	0.85 ± 0.004	$0.95 \pm 0.009^*$
	ViT	0.91 ± 0.003	$0.97 \pm 0.007^*$
	HDNet (Ours)	-	$\mathbf{0.98 \pm 0.001}$
Materials	ResNet	0.64 ± 0.03	$0.83 \pm 0.008^*$
	ViT	0.78 ± 0.01	$0.92 \pm 0.003^*$
	HDNet (Ours)	-	$\mathbf{0.94 \pm 0.002}$
Viewpoint	ResNet	0.63 ± 0.02	$0.72 \pm 0.009^*$
	ViT	0.77 ± 0.01	$0.83 \pm 0.001^*$
	HDNet (Ours)	-	$\mathbf{0.83 \pm 0.006}$

Table Sup1: **Adding scene context improves performance independent of architecture.** Following the design of HDNet shown in **Fig. 1(c)**, we modified standard architectures to have two streams—one operating on the target, and the other one on the contextual information. Representations for both streams are then concatenated and passed through a classification layer as shown in **Fig. 1(c)**. We train the standard one-stream and these modified two-stream architectures on HVD, and report the average Top-1 accuracy for all models . We also report error bars, which measures the variance in accuracies over categories. Both the ResNet and the ViT architectures lead to a large improvement in generalization for all semantic shifts when modified to leverage scene context. To ensure we study impact of context independent of data diversity, all models were trained on 4 domains, i.e., 80% transformational diversity and tested on the held out domain. Best performing model (HDNet) has been shown in boldface for all real-world transformations. A $*$ refers to statistically significant improvement in performance when using a two-stream architecture as compared to a one-stream architecture (two-sided t-test, $p < 0.05$).

Semantic Shift	Without Contrastive Loss	With Contrastive Loss
Viewpoint	0.79	**0.82**
Material	0.89	**0.94**
Lighting	0.98	**0.98**

Table Sup2: **Impact of removing contrastive loss**. We evaluate the contribution of the contrastive loss by training and testing HDNet on the HVD dataset with and without the contrastive loss. The contrastive loss results in an improvement across all three semantic shifts.

Semantic Shift	Full Context ($\sigma = 0$)	Less Context ($\sigma = 25$)	Least Context ($\sigma = 125$)
Lighting	$\mathbf{0.98 \pm 0.001}$	0.96 ± 0.001	0.94 ± 0.001
Material	$\mathbf{0.94 \pm 0.002}$	0.88 ± 0.01	0.83 ± 0.006
Viewpoint	$\mathbf{0.83 \pm 0.006}$	0.77 ± 0.01	0.76 ± 0.01

Table Sup3: **Blurring scene context worsens generalization performance**. We trained and tested HDNet with the scene context in HVD images blurred using a Gaussian blur. Here, σ is the standard deviation for the gaussian kernel applied to the image as a filter. Thus, blurring increases with σ. We applied three values for σ—0,25, and 125. For brevity, numbers less than 0.001 are reported as 0.001.

To further understand the role of contextual information on visual recognition, we conducted two additional experiments. Firstly, we evaluated the impact of reducing scene context information by blurring it using a Gaussian Blur. As shown in **Table. Sup3**, performance dropped consistently for all three transformations as contextual information is reduced. Secondly, we confirmed that the increase in performance is due to the addition of contextual information and not due to the two-stream architecture *per se* by training HDNet with both streams receiving only the target information. This removal of context led to a drop in performance, as reported in **Table. Sup4** (see **Sec. F** for details).

Besides results on the role of context presented in **Table. Sup1**, we present here two additional experiments evaluating the contribution of scene context on generalization. Firstly, we also evaluated the impact of blurring the scene context while keeping the target intact [47]. For each real-world transformation, we trained and tested models with increasing levels of Gaussian blurring applied to the scene context. These results are presented in Blurring was applied to the images in the form of a Gaussian kernel filter, with the kernel standard deviation (σ) set to 0, 25, or 125. The cropped image of the target object was passed to the second stream of the network without blurring. These results are reported in **Table Sup3**. As can be seen, there was a drop in performance as context blurred for all three real-world transformations.

Semantic Shift	Target only	Target and Context
Viewpoint	0.77	**0.82**
Material	0.85	**0.94**
Lighting	0.97	**0.98**

Table Sup4: **Training a two-stream HDNet with only target information**. As a third control for confirming the role of context, we train HDNet where both streams are passed just the target object. Thus, it is forced to learn without scene context. This results in a drop in performance for all semantic shifts, providing further evidence in support of the utility of scene context.

Secondly, we train HDNet such that both streams are trained with the target object. Thus, this modified version is forced to learn without scene context. These results are shown in **Table. Sup4**. For all semantic shifts, forcing HDNet to learn with only the target results in a drop in accuracy. This provides further evidence supporting the utility of scene context in enabling generalization.

G Additional experiments with HDNet and contrastive loss

We evaluate the contribution of the contrastive loss by training variations of HDNet on HVD with and without the contrastive loss as shown in Eq. Sup2. These numbers are reported in **Table Sup2**. As can be seen, adding a contrastive loss improves performance for all three semantic shifts, providing evidence for its utility.

H Additional experiments with a larger, less controlled ScanNet test set.

We extend the generalization to real-world results presented in the main paper by reporting these numbers on a larger test set created by annotating additional images from ScanNet. As ScanNet

Test Dataset	ResNet [73]	ViT [60]	AND Mask [28]	CAD [34]	COR AL [29]	ERM [32]	IRM [30]	MTL [61]	Self Reg [31]	VREx [33]	HDNet (ours)
ScanNet	0.35	0.29	0.43	0.40	0.42	0.48	0.46	0.46	0.53	0.42	**0.61**

Table Sup5: **Human visual diet improves generalization to larger real world dataset as well**. We curated a larger subset of ScanNet images, allowing more complex real world scenarios like blurry images, clutter and occlusions. We report the capability of models to generalize from synthetic HVD images to this more complex subset of ScanNet. HDNet leveraging human-like visual-diet outperforms all baselines on this more complex dataset as well.

was created by shooting video footage of 3D scenes, many frames can be blurry. In the original, smaller test-set such blurry frames were removed to ensure a higher quality test set. However, here we also include additional images with lower fidelity to report numbers on a larger test set. These numbers are reported in **Table. Sup5**. The trend is consistent with results reported on a smaller, more controlled subset in the main paper—HDNet outperforms all other benchmarks by a large margin. As expected, including these images in the test set results in a drop in accuracy across all methods. All models were trained on synthetic images from HVD and were tested on a test set of natural images from ScanNet.

I Hyperparameters

HDNet: As our model builds on top of CRTNet [63] 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: We used the code from Gulrajani et al. [74] to train and test domain generalization 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 top performing algorithms as the baselines reported in the paper.

FasterRCNN: We used the code from Bomatter et al. [63] 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 repository.

J Experimental Details

HDNet was compared against several baselines presented below. All models were trained on NVIDIA Tesla V100 16G GPUs. Optimal hyper-parameters for benchmarks were identified using random search, and all hyper-parameters are available in the supplement in **Sec. I**.

J.1 Baseline Approaches

We compared the impact of a human-like visual diet with a diverse set of alternative approaches popular in machine learning. This includes:

2D feed-forward object recognition networks: Previous works have tested popular object recognition models in generalization tests [75, 76]. We include the same popular architectures ranging from 2D-ConvNets to transformers: DenseNet [77], ResNet [73], and ViT [60]. These models do not use context, and take the target object patch I_t as input.

Domain generalization methods: We also compare HDNet to an array of state-of-the-art domain generalization methods (**Table 1**). These methods also use only the target object, and do not use contextual information.

Context-aware recognition models: To compare against models which use scene context, we include CRTNet [63] 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 compare HDNet with a Faster R-CNN [62] model modified to perform recognition by replacing the region proposal network with the ground truth location of the target object.

Billion-Scale self and semi supervised architectures: We presented results with a suite of modern approaches trained on 1000-fold more data to emphasize the importance of data quality over sheer dataset size. These included—Dino V2, ResNet50 SWSL, ResNet18 SWSL, 32x4d SWSL, ResNext101 32x16d SWSL, and ResNext50 32x4d SWSL.

J.2 Evaluation of computational models

Performance for all models is evaluated as the Top-1 classification accuracy. Error bars reported on all figures refer to the variance of per-class accuracies of different models. For statistical testing, p-values were calculated using a two-sample paired t-test on the per-category accuracies for different models. The t-test checks for the null hypothesis that these two independent samples have identical average (expected) values. For ScanNet, a t-test is not optimal due to the smaller number of samples, and thus a Wilcoxon rank-sum test was employed for hypothesis testing as suggested in past works [78, 79]. All statistical testing was conducting using the python package *scipy*, and the threshold for statistical significance was set at 0.05.