COLORSENSE: A Comparative Study on Color Vision in Machine Visual Recognition

Ming-Chang Chiu, Yingfei Wang, Derek Eui Gyu Kim
University of Southern California
{mingchac,yingfei,euigyuki}@usc.edu

Pin-Yu Chen
IBM Research
pinyu@ibm.com

Xuezhe Ma
University of Southern California
xuezhema@usc.edu

Abstract

Color vision is essential for human visual perception, but its impact on machine perception is still largely unexplored. There has been an intensified demand for understanding its role in machine perception for safety-critical tasks such as assistive driving and surgery but lacking suitable datasets. To fill this gap, we curate multi-purpose datasets, COLORSENSE*, by collecting* ***110,000 non-trivial human annotations*** *of foreground and background color labels from popular visual recognition benchmarks. To investigate the impact of color vision on machine perception, we assign each image a color discrimination level based on its dominant foreground and background colors, and use it to study the impact of color vision on machine perception. We validate the use of our datasets by demonstrating that color discrimination level has a dominating effect on the performance of mainstream machine perception models. Specifically, we examine the perception ability of machine vision by considering key factors such as model architecture, model size, training data and task complexity. Furthermore, to investigate how color and environmental factors affect the robustness of visual recognition in machine perception, we integrate our* COLORSENSE *datasets with image corruptions and perform a more comprehensive visual perception evaluation. We develop a dual-objective framework to jointly analyze the impact of color vision and image corruption on machine perception. Our findings suggest that* object recognition *tasks such as* classification *and* localization *are susceptible to color vision bias, especially for high-stakes cases such as vehicle classes, and advanced training techniques such as adding data augmentation only give marginal improvement. Our study highlights the need for new evaluation approaches and training techniques that can achieve human-aligned machine perception and improve the performance of machine perception models in real-world applications. Lastly, we showcase potential applications of* COLORSENSE *such as studying spurious correlations.*

1. Introduction

The development of deep neural networks (DNN) has been deeply rooted in the inspiration drawn from neuroscience and cognitive science [46]. For instance, convolutional neural networks (CNN) [19, 31] have enjoyed great success in various tasks in computer vision by modeling the human visual cortex system. More recently, Transformer [53] has become a widely adopted architecture, which models the human attention mechanism that allows for better contextual understanding. It has since extended its success beyond natural language processing to computer vision [15]. Some recent studies explore behavioral aspects of DNN models for object recognition [17, 24, 52]. However, there exist other important parts of the primates' visual pathway that remain underexplored, such as *color vision* [28, 45], a neuroscience domain that studies how color affects human visual perception. With such a nomination, we explore beyond the regime of using total accuracy as the sole objective and curate new datasets to study machine visual recognition as a proxy task to connect the *color vision* aspects of human visual perception with those of machine vision.

Color is a fundamental perception that we often take for granted as humans. It is only when our color vision is hindered, as in cases of colorblindness, that we realize its true significance. Similarly, color plays a role in machine vision [39] and has a crucial impact on safety-critical downstream visual recognition tasks such as self-driving, skin disease detection, or robotic-assisted surgery.

In this paper, we aim to establish a first framework to study how color signals can affect current machine visual

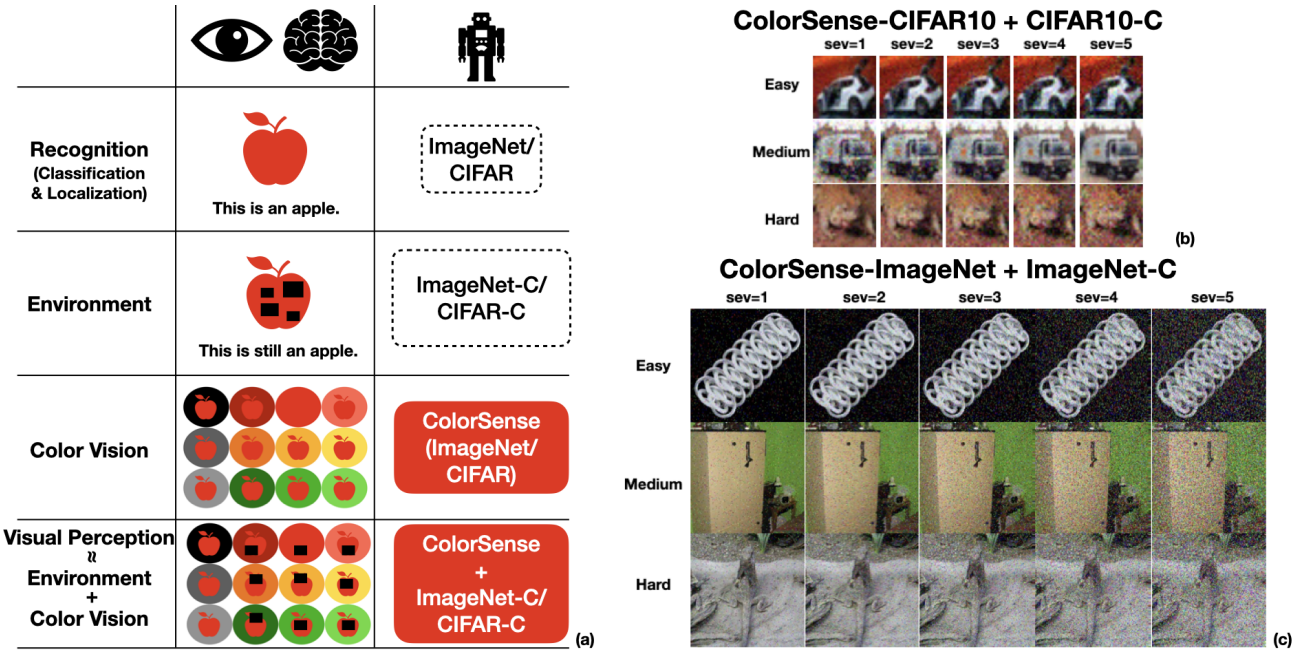


Figure 1. **Mapping of our ColorSense datasets to human visual aspects and examples of our ColorSense integrated with corrupted images (ImageNet and CIFAR10).** (a) While common benchmarks can evaluate machine visual recognition (overall accuracy) and simulate corrupted images, ColorSense evaluates the color vision capability. (b-c) Examples of our ColorSense integrated with CIFAR10-C/ImageNet-C datasets. Rows mean color discrimination levels.

perception models. To do so, we manually separate ImageNet into eight groups based on the dominant foreground and background color of each image following similar guidelines to [10]. Furthermore, to study how neighboring colors help machines discriminate objects, we consult color professionals and the established standard [1] to categorize the foreground/background color pairs into color discrimination groups (details in §3), namely {Hard, Medium, Easy}. We call this dataset ColorSense, designed to quantify the sense of color for machine vision. Similarly, on CIFAR, we also provide dominant foreground color attributes, and along with CIFAR-Background (CIFAR10-B) [10] (Fig. 2(b)), to create ColorSense-CIFAR. We mainly focus on ColorSense-ImageNet in this work.

As a benefit, ColorSense can also be seamlessly integrated with out-of-distribution (OOD) datasets like ImageNet-C [22] (dubbed ColorSense-ImageNet-C) as a complement to broaden the analysis of color vision to include simulated environmental factors such as *noisy* data shifts. For human vision, it is well-established that the primate visual system can effortlessly and robustly discriminate between various visual objects, regardless of the wide range of images that each object can produce during natural vision [14, 16, 26, 36, 44, 47, 48, 51, 63], and the robustness should extend to different surroundings or added noise [28, 37, 45]. Fig. 1 illustrates where ColorSense fits in bridging the evaluation of visual perception between humans and machines.

To gain a comprehensive understanding of color vision in machine vision, we develop a framework for testing ColorSense in conjunction with corrupted datasets across different model sizes and architectures, which shows the diverse capabilities of ColorSense. This framework provides new insights into model behaviors under different environmental and color discrimination conditions. Our study is the first step toward understanding color vision in deep learning, and we believe ColorSense will facilitate benchmarking and evaluation efforts in different ways. Lastly, we showcase other uses of our dataset: we perform case studies on a safety-critical task to demonstrate the importance of color vision, quantify model robustness and study fairness using color subgroups from ColorSense.

To summarize, our contributions are fivefold:

- We curate ColorSense dataset with 110,000 human annotations to study how color vision affects computer vision models on various visual recognition tasks.
- We integrate our dataset and ImageNet-C to study the robustness of machine color vision in natural and simulated noisy environments.
- We conduct comprehensive analyses on factors like model size, architecture, training procedures, and two vision tasks.
- We propose generic metrics for the color vision effect and model robustness.
- We show ColorSense's diverse uses, such as studying spurious correlation.

2. Bridging Human Vision and Machine Vision

2.1. Human Vision Test

We, as humans, rely on stimuli from light that reflect on objects to our eyes to perceive objects. More importantly, our retinal cone cells respond differently to wavelength in a diverse spectrum of light [28, 45]. For instance, three different types of cone cells respond predominately to red, green, and blue. This *color vision* mechanism helps us discriminate objects under different *luminance, contrast, environments*, etc., and is an essential part of why our vision can adapt quickly [20, 28, 45] and thus become more robust [14, 16, 26, 36, 44, 47, 48, 51, 63]. More concretely, the opponent-process theory, a theory that different wavelengths would oppress one type of cell [28, 45], explains how psychological perception would affect our way of discriminating objects [28, 45]. Such phenomena have important real-life applications, for example, a red apple would be more clearly identified when placed on a green background (Fig. 1) than on an orange background; army soldiers would wear green and brown uniforms instead of white for better camouflage. Note that as color constancy remains the same for machines (same pixel values), color vision and color discrimination represent the same concept [4] in our work and will be used interchangeably.

To measure human visual perception ability, a common routine is to perform the *Snellen's test*, where subjects need to discriminate black alphabets in a white background from a certain distance. This test measures our *visual acuity*, the ability to distinguish shapes and the details of objects [40, 43]. However, a 20/20 vision does not indicate perfect vision [27, 40, 43], and ophthalmologists have proposed the *contrast sensitivity test* and some color vision tests. The former measures the ability to discriminate finer and finer increments of light versus dark [21], and the latter includes *color plate test, hue test, anomaloscope test* [2] to measure the accuracy of color vision.

2.2. Machine Vision Test

Unlike various early visual perception tests for human vision, machine vision often goes directly to recognition tasks. Although DNN models reach human-level performances on diverse visual recognition tasks [54–56], they are not always robust [9, 18, 22], which means there are *notable gaps to true human-aligned vision*. Hence, instead of revolving around performances on generalization errors, we take a step back and ask: *What are the fundamental aspects that current studies overlooked?* Inspired by the neuroscientific literature, our answer is to study *how machine color vision affects machine visual recognition*, as an important *first step* to bridge the gap. Prior works have touched on the related *contrast sensitivity* concept [5, 32] and Olah et al. [39] visualizes early CNN layers to hint color's role. But due to the lack of proper datasets, they *did not* evaluate DNNs' *color vision* behavior that affects visual recognition.

CD Group	#pairs	ImageNet	CIFAR10
Hard	7	9121	1690
Medium	12	15835	3433
Easy	9	18503	4126
Others	8	6541	751

Table 1. Statistics of COLORSENSE by color discrimination group.

3. Benchmarking for Color Vision

In the section, we introduce our COLORSENSE-IMAGENET, a dataset that are based on the validation sets of ImageNet [13]. On ImageNet validation set, we label 50,000 images' dominant colors in both the foreground and background. To enable color vision study in different setups, we also provide 10,000 color labels on CIFAR10 [30], forming COLORSENSE-CIFAR. With our labeling efforts, we can create color discrimination groups for measuring color discrimination ability in computer vision and visual recognition. Note that in this work, we focus on the evaluations on ImageNet.

3.1. Labeling Process

1. We label the perceived color that has the most coverage on an object. For example, in Fig. 2 (c-*top*), the car is almost red everywhere so we label foreground as "red."
2. When two or three colors have significant coverage, we choose the color that has more coverage. We label the example in Fig. 2 (c-*mid*) as "brown", though the black track is also significant.
3. When the perceived color does not belong to our categories, or when multiple colors appear on the object and none is larger than the rest, we put it in the "others" category (Fig. 2 (c-*bottom*)).

3.2. Defining Color Discrimination Groups

We follow the Web Content Accessibility Guidelines [1] and consult experts from the School of Fine Arts to assign all possible color pairs into three color discrimination (CD) groups, {HARD, MEDIUM, ESAY}. The grouping is primarily based on the relative luminance of each color pair (Fig. 2). The HARD groups consist of images with the same foreground and background colors. The EASY group includes color pairs with a score ranked in the top one-third of all pairs, and the rest goes into MEDIUM. Tab. 1 presents the statistics for the number of color pairs in each CD group and the number of test images in each group for COLORSENSE datasets. In the following sections, we mainly focus on the evaluations on COLORSENSE-IMAGENET and touch lightly on COLORSENSE-CIFAR10 due to the space limit. Details of the color pairs for each color discrimination group are provided in the Appendix.

4. Evaluation: Image Classification

4.1. Fundamental Questions

Does color vision affect DNN models as humans? While DNNs are artificial models of neurons in the brain,

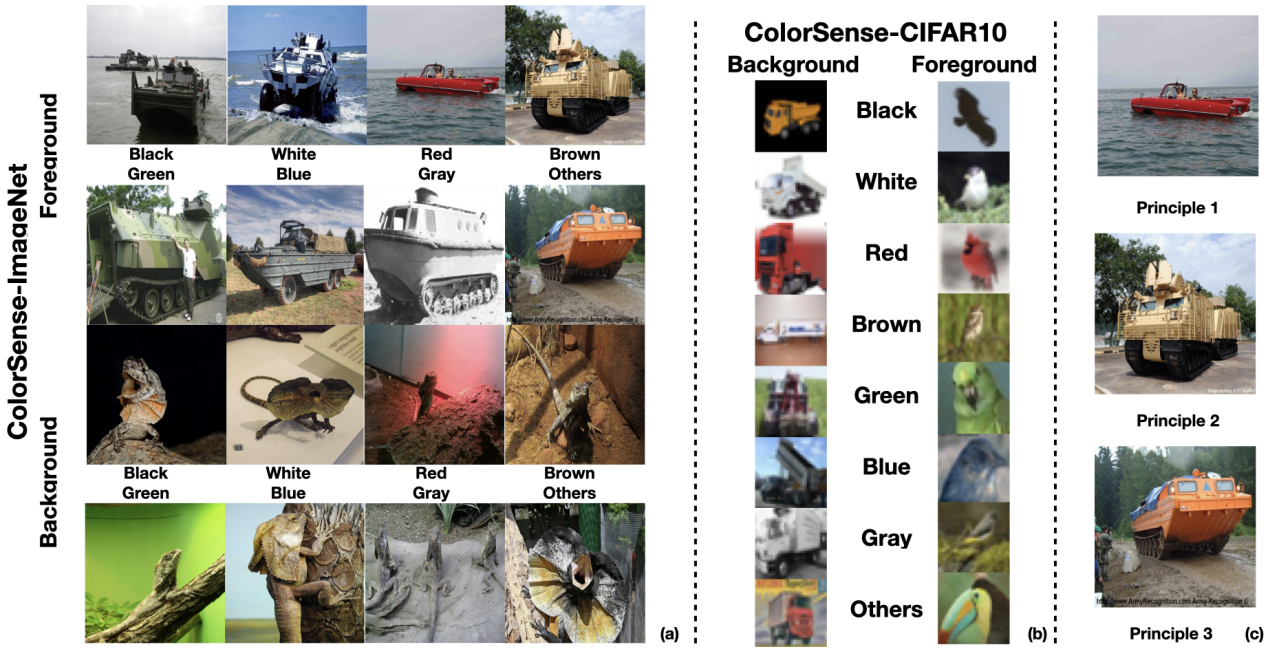


Figure 2. **Examples of our COLORSENSE datasets and labeling principles.** (a) The top two rows are COLORSENSE-IMAGENET-FOREGROUND and the bottom two rows are COLORSENSE-IMAGENET-BACKGROUND. In conjunction, we can define color discrimination groups (§ 3). (b) The left and right column are COLORSENSE-CIFAR10-BACKGROUND/FOREGROUND. (c) Examples of the three labeling principles. See § 3.1 for full labeling details.

some of them are designed to mimic the structure and function of the visual cortex [31]. As such, it is possible that these models possess some form of color discrimination, similar to how humans perceive color. This hypothesis raises the question of whether color vision affects DNN models and how this could impact their performance in various applications. Further research is needed to explore the relationship between color vision and DNN models and to uncover any potential benefits or limitations. Most humans are robust to color vision groups, but the decisions vary in response time [14, 16, 17, 26, 36, 44, 47, 48, 51, 63], whereas for machines, the distinction could be reflected in the form of performance difference. We also conduct a small scale human study to verify it in the Appendix.

Does architecture matter? The choice of deep learning architecture can have a significant impact on the performance and efficiency of a model [19, 33–35, 50]. CNNs, such as ResNets, have been the go-to architecture for computer vision tasks, thanks to their ability to learn spatial features from images. However, the emergence of newer architectures, such as the ViTs, has challenged this dominance. Unlike traditional CNNs, ViT uses self-attention mechanisms to capture global image features, allowing it to outperform CNNs on some tasks and claim less inductive bias. The debate over whether architecture matters and which one to choose, continues to be an active area of research and development. We compare diverse architectures to answer this question and hope to shed light on this area in terms of color vision.

Does model size matter? The relationship between model size and robustness is an intriguing topic of research in the field of deep learning. Hendrycks et al. [23] showed that larger models tend to be more robust to added noises. This finding is particularly interesting since larger DNNs do resemble the number of neurons in the human brain more closely. Furthermore, human vision is known to be robust to color variations, which raises the question of whether larger DNNs are also more robust to color variations. Exploring the relationship between model size, robustness, and color variation could lead to useful insights into the functioning of DNNs and their potential applications. We compare (1) architectures of the same family and (2) models of similar sizes but different architectures to answer this question in our problem setup.

4.2. Findings on COLORSENSE-IMAGENET

To answer the three fundamental questions in Sec. 4.1, we conduct experiments with a diverse set of models trained on ImageNet (model weights are fixed and provided by PyTorch). To account for the wide range of application setups, we select both classic and modern CNNs such as Resnet [19], Resnext [60], Convnext [35] and MobileNetV2 [50], and vision transformers such as ViT [15] and Swin Transformers [33, 34].

DNNs are deeply affected by color vision. We observe a consistent performance increase from the HARD CD group to the EASY group (Fig. 3) across all the aforementioned architectures. We perform *Paired t-Tests* between the CD groups across all models. We pool the HARD group accu-

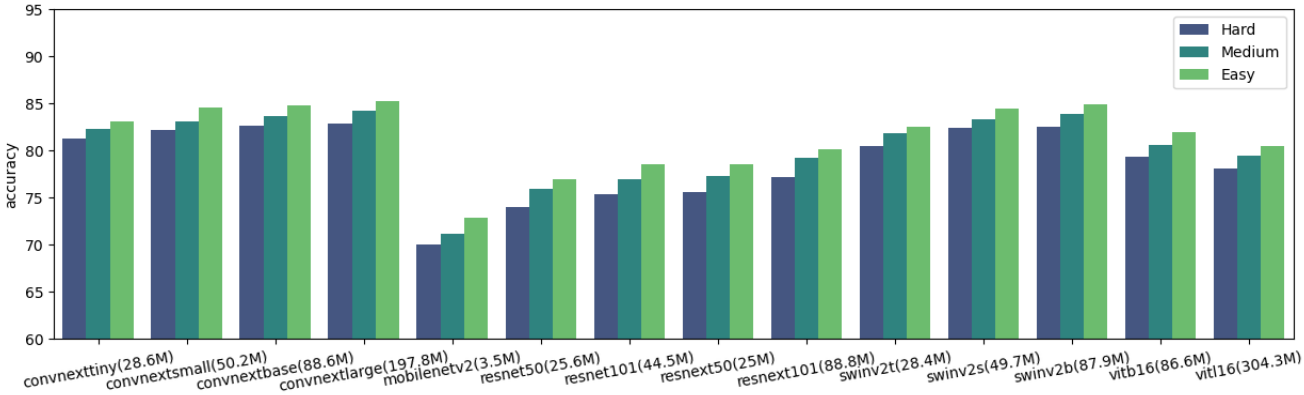


Figure 3. **Testing models with different architectures and model sizes (ImageNet).** We observe similar trends in the three CD groups (Hard, Medium, Easy) across all models, showing the effect of color vision is universal across architectures and model sizes.

racies from all models and EASY group accuracies from all models and then performed the statistical tests on these two sets of data. The p-value for HARD-MEDIUM is $1.39e$-6, HARD-EASY is $8.34e$-7, and MEDIUM-EASY is $9.87e$-6, which are all statistically significant. We also perform a *Page's Trend Test* for increasing means ($H_0 : \mu_{Hard} = \mu_{Medium} = \mu_{easy}, H_A : \mu_{Hard} < \mu_{Medium} < \mu_{Easy}$) and the p-value is $5.95e$-7, which is also statistically significant. These results suggest that the DNNs do have color vision ability-related issues, and color discrimination abilities should be taken into account in the future when designing and evaluating computer vision models. Second, it suggests that there is room to improve the performance of DNNs on color-based tasks by developing models that can better discriminate between colors. These implications point out a deficiency of DNNs models and shed light on an area of focus for the improvement of next-generation computer vision applications.

Model size and architecture do not add obvious robustness to color vision. For each model architecture, we evaluate its multiple variations based on size. To our surprise, Fig. 3 shows that the performance gaps between the EASY and HARD groups do not obviously decrease as model size increases, suggesting the impact of the color discrimination group remains almost unchanged. In addition, we compare model architectures while controlling for model size; for example, models with 88M parameters (`Convnext_base`, `Resnext101`, `SwinV2_b`, `ViT16_b`) show similar performance gaps. While model size and architecture are essential factors to consider when designing and evaluating DNNs for various computer vision tasks, they may matter less when it comes to color vision discrimination — we observe an upward trend in performance towards the EASY group regardless of the architecture or size of the model. These suggest that human-aligned color discrimination ability may not be inherent in DNNs, and developing specialized models or techniques to address color vision discrimination may be necessary.

Absolute Gap	Conv.	Mob.	Mob(adv)	Res.	Res(adv)	Resx.	Resx(adv)	SwinV2	ViT
COLORSENSE	2.17	2.86	2.62	3.12	2.21	2.97	2.56	2.18	2.48
Land vehicle-only	3.40	4.35	4.87	6.59	4.27	4.63	5.87	4.31	6.93
COLORSENSE-C	4.59	3.37	n/a	4.40	n/a	4.46	n/a	4.40	4.64

Table 2. Average absolute gap by architecture type and dataset.

4.3. Ablation: Advanced Setups

We have identified the role of color vision in DNNs. To comprehensively understand its relationship with DNNs, we must consider additional factors across the entire deep learning pipeline, encompassing advanced training methods, high-stakes applications, OOD scenarios like grayscale images, and environmental influences. These aspects will be discussed in this section. Furthermore, we take into account the influence of training data and conduct zero-shot classification using foundation model like CLIPs.

Advanced training setups do not add significant performance robustness to color vision groups. Recent works have shown that specific training recipes, such as longer training epochs and data augmentations, can enhance model performances [11, 59, 62]. Data augmentation has long been regarded as the most effective approach for achieving robustness [57, 61]. However, our results suggest no significant changes in overall robustness to color discrimination groups (Fig. 4), as measured by *absolute gap* (AG), $AG = |acc_{max} - acc_{min}|$, where $acc_{max/min}$ is the maximum/minimum accuracy of the three groups (Tab. 2). Advanced training only reduces AG by an average of 0.58% in accuracy. This bolsters our belief that the color vision phenomenon exists in machine vision.

Less bias on grayscale images. To eliminate confounding factors, we perform analysis on a control group: grayscaled ImageNet. We still observe lower performance in harder CD groups. To compare the effect, we quantified the absolute gap between the HARD and EASY groups for color and grayscale images, respectively. Fig.5-*right* shows that color vision effect is more pronounced in colored images, reaffirming our primary finding.

Environmental factors play a role. A benefit of COLORSENSE is the easy incorporation with OOD datasets

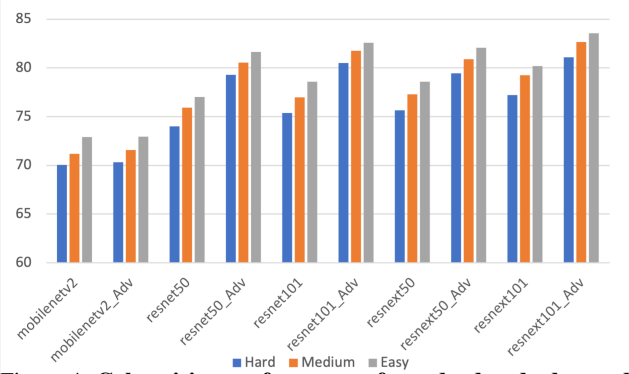


Figure 4. **Color vision performance of standard and advanced procedure-trained models.** Advanced training only makes models slightly more robust to color vision effect (Tab. 2).

like ImageNet-C to task models with added environmental factors and noises. We observe a consistent trend that color vision plays an important role in all types of corruption. In Fig. 6, we show three models, `Convnext_s`, `Swin_b`, `SwinV2_s` of similar accuracies on ImageNet, 83.61, 83.58, and 83.71, respectively (less than 0.15 difference) for comparisons. Even though the generalization performances are almost the same, for each corruption type, the color vision performance is different. For instance, `Swin` & `SwinV2`'s color discrimination capabilities are better than `Convnext_s` for all corruptions but *jpeg compression*, and `SwinV2_s` is slightly better than `Swin_b` on *digital* corruption such as *pixelate* but worse on *noise* corruptions. These findings have implications and insights for decisions such as *which model we should choose?* If we expect noisy environments for some custom application, we should choose `SwinV2_s` model among the three.

Also, we conduct case studies on specific corruptions, such as Gaussian noise (Fig. 6 (*bottom-right*)) due to its commonality in image processing. We observe that CNNs generally have less obvious gaps in MEDIUM and EASY than ViTs. This analysis reveals the unique advantage of CNNs over ViTs. Similar analyses can be done for other corruptions. One can easily use our analysis framework to assess different models and provide guidance on model selection in more realistic settings, considering the effect of color vision.

Zero-shot Foundation Models behave similarly. We include CLIP zero-shot classification results (Fig. 7-*right*) for ViT-B32 and ViTB14 with two different training data: Laion400M and Laion2B. We observe a general trend: color vision is still in effect.

Changing training data does not alter the color discrimination phenomenon. We also include the results of ViTs pretrained with ImageNet21k. We observe that on ViTs, the color vision effect is present (Fig. 7-*left*).

Together with advanced training and CLIP zero-shot results, the results suggest that *training procedures do not alter the phenomenon.*

High-stakes examples have larger gaps. Color vision is important for humans to perform high-stakes tasks such as driving or cardiovascular surgery. Prior work has shown potential lighting changes can lead to dangerous erroneous behavior [42]. With the popularity of self-driving cars and the previously established knowledge about color vision in machine vision, we apply our analysis specifically to the vehicle classes in ImageNet. Ideally, we want models to be the least affected by color vision on the car-related subset. *To our surprise*, we observe models generally have *even larger gaps* (Tab. 2 & Fig. 8-*left*) between CD groups. It suggests vehicle classes are affected by color vision even more, and surprisingly, some larger models, such as `ViT_l` is affected more than `ViT_b`. Advanced training (Tab. 2) only reduces the color vision gap for `Resnets` but increases for `Resnexts` and `MobileNetV2`. Therefore, our result suggests that models like `MobileNetV2` may not be suitable to be deployed for self-driving systems. Our evaluations call for the need for further color vision studies for high-stakes tasks.

Quantifying overall model ability to CD effect. To jointly consider the CD effect and performance on corrupted dataset, we propose two metrics: *scaled color vision robustness* (sCVR(↑)), $\frac{acc_{tot}}{\sigma_w}$, where acc_{tot} is the total accuracy on ImageNet-C and $\sigma_w = \sqrt{\frac{\sum_{i=Easy}^{Hard} w_i(acc_i - \bar{acc})^2}{\sum_{i=Easy}^{Hard} w_i}}$, is the weighted standard deviation across CD groups. Similarly, we define *scaled corruption robustness* (sCR(↑)), $\frac{acc_{tot}}{\sigma_s}$, where $\sigma_s = \sqrt{\frac{\sum_{s=1}^{5}(acc_s - \bar{acc})^2}{\sum_{s=1}^{5}}}$ (across five severity levels). Fig. 9-*left* shows the sCVR and sCR of each model. Per our metrics, `Convnext_l` is the overall best. We also tried the $\frac{acc_{tot}}{AG}$ as a metric, and the results were similar. We observe that most models have better sCVR as their size grows. However, for ViTs, as also observed in the land vehicle classes, there is an inverse correlation to model size, which is another anomalous model behavior our dataset reveals. We discuss the practicality of our metrics in the Appendix.

4.4. Results on COLORSENSE-CIFAR10

We apply Resnet, Densenet [25] and ViT as backbones to evaluate on COLORSENSE-CIFAR10. The training details of these models are provided in the Appendix. In general, we observe a similar trend of having better performance in easier groups as on ImageNet. Fig. 8-*right* summarizes the performances of both pre-trained models (pre-{model}) and those trained from scratch. Interestingly, not all models are affected consistently by color vision, and pre-training on ImageNet also does not help reduce AG consistently. Though the results on CIFAR10 and ImageNet are slightly different, we believe the ImageNet results and findings are more representative of real-world scenarios.

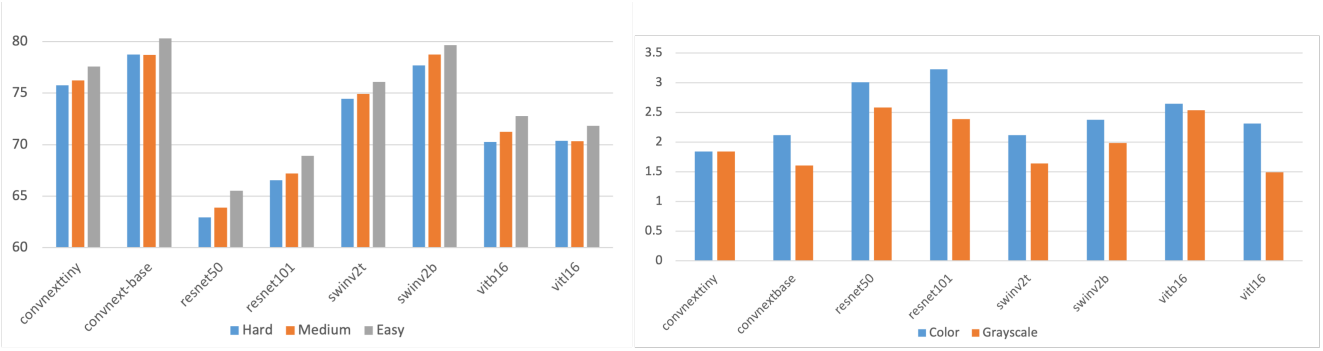


Figure 5. *Left:* **Grayscale ImageNet results (%).** We still observe slight color vision effect between the CD groups. *Right:* **Absolute gap (AG) comparisons.** The color vision effect is more pronounced in color image, supporting our primary finding.

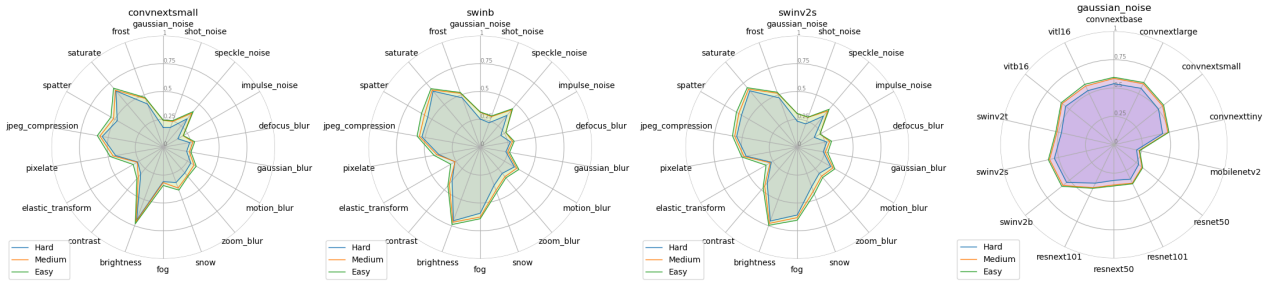


Figure 6. **Color discrimination capabilities with added noises by models (*top row & bottom-left*) and by Gaussian noise (*bottom-right*).** Models of the same total performances can have very different color vision behaviors in the presence of different corruptions. *Bottom-right*: CNNs see less obvious gaps between MEDIUM and EASY group than transformers when presented Gaussian noise. These phenomena show that each architecture is built differently and suitable for different scenarios, and our analysis framework can serve as a cookbook for downstream applications.

CD Group	IoU-30	IoU-50	IoU-70
Easy	75.31	73.12	50.34
Medium	74.48	72.77	50.01
Hard	72.94	70.80	48.97

Table 3. Object localization accuracy.

5. Other COLORSENSE Usages: Object Localization and Spurious Correlation

Our COLORSENSE has enormous potential for a wide range of applications in computer vision. We already show its usefulness for visual perception tasks such as image classification, where color plays a critical role in recognizing objects in an image. We conduct another key visual recognition task, object localization (OL). Other than that, we can study color as a spurious correlation to analyze subgroup performance. In general, COLORSENSE is a valuable tool for any task that studies color discrimination in computer vision.

Object localization is similar to classification. We apply the same analysis procedure in Sec. 4.2 to object localization [7]. We report top-1 localization accuracy with three different IoU thresholds (30, 50, 70). We summarize the results on COLORSENESE-IMAGENET in Tab. 3. We observe similar performance differences as in classification task, where the model tends to perform better as the CD group becomes easier. And this trend happens across all three different IoU thresholds. This again bolsters our finding of the color vision effect in machine vision.

High-stakes examples in object localization. Similar to classification, we conduct a case study on land vehicle classes for object localization due to their resemblance to a self-driving setup. We also observe larger AG between CD groups, meaning an amplified effect of color vision. Fig. 9-*mid* shows the IoU-30 result and we observe similar trends using other thresholds. We emphasize the worst-case situation in high-stakes tasks. For the HARD group, we observe worse performances on vehicle classes than all classes in all three thresholds, suggesting higher safety risks in those conditions and the necessity of risk mitigation. We believe our findings can be extended to other self-driving studies.

Background and foreground color as spurious correlations. Another use of our dataset is to leverage the background/foreground labels as spurious contextual correlations. Though this is not the main focus of this work, we conduct a similar analysis as in [10] to showcase this application of our COLORSENSE. For example, on `ResNet50`, we observe "subgroup degradation," uneven performances across subgroups, for both foreground and background. For instance, `ResNet50` tends to bias in favor of green objects and objects in a green background (Fig. 9-right). Also, we observe similar trends across all models, showing the universality of this bias. The complete plot for all models is shown in Appendix.

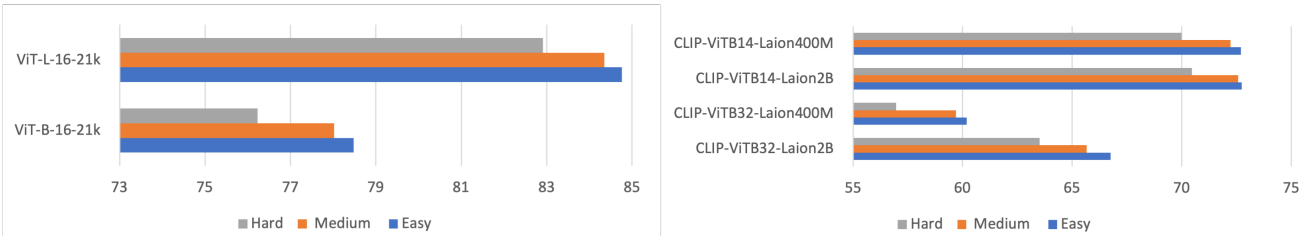


Figure 7. **Color vision in effect no matter how models are pre-trained.** *Left:* ViT accuracies (%) pre-trained on ImageNet-21k. *Right:* CLIP 0-shot performances (%) with different training data size.

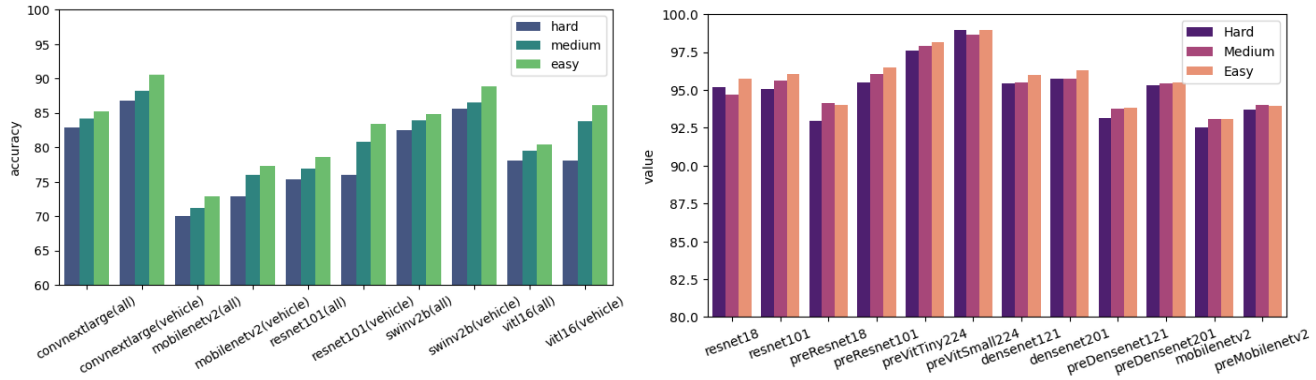


Figure 8. **Comparisons of all vs. land vehicle classes on ImageNet (*left*) & CIFAR10 performances (*right*).** *Left:* overall larger gaps between HARD and EASY group on vehicle classes suggest potential safety concerns when deploying such models on the road. *Right:* the trend of better performances on EASY group persists on CIFAR10.

6. Discussion

Our work presents the development and demonstrates the utilities of our curated COLORSENSE dataset, which is a versatile tool for (1) investigating the impact of color vision on various visual recognition tasks, (2) featuring compatibility with OOD environmental noise datasets, and (3) controlling for spurious correlations. We argue that current literature on machine vision evaluations does not fully explore the role of color vision in object recognition, and show that our framework provides a more comprehensive approach to evaluating machine visual recognition.

Our work demonstrates that current machine vision is susceptible to color vision bias, regardless of DNN architecture or model size. We provide evidence that machine vision is affected by color vision in the form of performance reduction in harder CD groups, which suggests that more innovations are needed to improve model robustness to color to align with human vision. We also show that advanced training procedures for robustness enhancement do not mitigate the color vision effect by much. Further, in the high-stakes use cases studied, the color vision effect may be more obvious. Lastly, CNNs such as `Convext` can be as powerful as transformers, if not better. Our analysis framework provides new guidance for model choice, given the expected application environment and computational budget.

The exploration of machine-human alignment stands as a highly coveted focus in research [3]. It is anticipated that advancements in our understanding of human vision [12, 37, 49] will pave the way for notable breakthroughs in deep learning, specifically addressing the color vision bias in models. Our COLORSENSE dataset is positioned as a versatile and valuable resource for the comprehensive assessment of newly proposed models and training methods [6, 8]. This dataset can not only be used to facilitate the reduction of model bias but also contribute to heightened awareness of safety and fairness concerns, thereby enriching the computer vision community.

References

[1] Web content accessibility guidelines (WCAG) 2.0, 2008. 2, 3

[2] Testing for color blindness, 2019. 3

[3] Our approach to alignment research, 2022. 8

[4] Alicia B Abrams, James M Hillis, and David H Brainard. The relation between color discrimination and color constancy: when is optimal adaptation task dependent? *Neural Computation*, 19(10):2610–2637, 2007. 3

[5] Arash Akbarinia, Yaniv Morgenstern, and Karl R Gegenfurtner. Contrast sensitivity function in deep networks. *bioRxiv*, pages 2023–01, 2023. 3

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 8

[7] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 612–628. Springer, 2022. 7

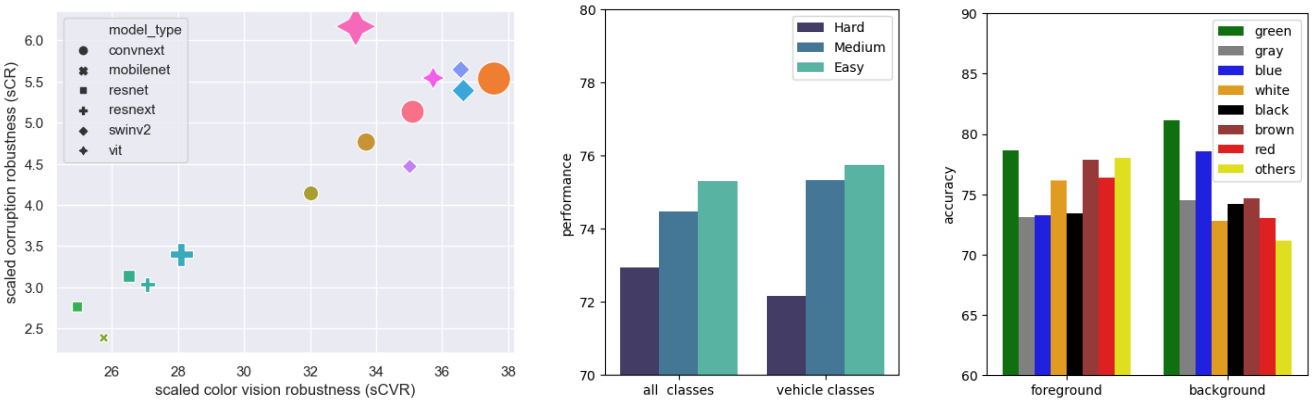


Figure 9. *Left:* overall quantified color vision robustness (symbol size proportional to model size). *Mid:* object localization performance on ImageNet (threshold=IoU-30) — vehicle classes are more affected by color vision effect. *Right:* background and foreground colors as spurious correlations; green objects/background are favored by `Resnet50`.

[8] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical programming*, 88:411–424, 2000. 8

[9] Pin-Yu Chen and Cho-Jui Hsieh. *Adversarial Robustness for Machine Learning*. Elsevier, 2023. 3

[10] Ming-Chang Chiu, Pin-Yu Chen, and Xuezhe Ma. Better may not be fairer: A study on subgroup discrepancy in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 7

[11] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *ArXiv*, abs/1805.09501, 2018. 5

[12] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020. 8

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

[14] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73 (3):415–434, 2012. 2, 3, 4

[15] 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. 1, 4

[16] Shimon Edelman et al. *Representation and recognition in vision*. MIT press, 1999. 2, 3, 4

[17] Robert Geirhos, Kristof Meding, and Felix Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *ArXiv*, abs/2006.16736, 2020. 1, 4

[18] I. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 3

[19] 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. 1, 4

[20] David Heeger. Perception lecture notes: Light/dark adaptation, 2006. 3

[21] Gary Heiting. Contrast sensitivity testing, 2019. 3

[22] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2, 3

[23] 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, 2019. 4

[24] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020. 1

[25] 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. 6

[26] Helene Intraub. Presentation rate and the representation of briefly glimpsed pictures in memory. *Journal of Experimental Psychology: Human Learning and Memory*, 6(1):1, 1980. 2, 3, 4

[27] Gregory J. Pamel and Dan Reinstein. Contrast sensitivity metrics extend beyond measure of vision. 2015. 3

[28] G.H. Jacobs. *Encyclopedia of Biological Chemistry (Second Edition)*. Elsevier, 2013. 1, 2, 3

[29] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2019. 1

[30] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 3

[31] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and

Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1, 4

[32] Qiang Li, Alex Gomez-Villa, Marcelo Bertalmío, and Jesús Malo. Contrast sensitivity functions in autoencoders. *Journal of Vision*, 22(6):8–8, 2022. 3

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[34] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 4

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 4

[36] Nikos K Logothetis and David L Sheinberg. Visual object recognition. *Annual review of neuroscience*, 19(1):577–621, 1996. 2, 3, 4

[37] Velitchko Manahilov, Julie Calvert, and William A Simpson. Temporal properties of the visual responses to luminance and contrast modulated noise. *Vision research*, 43(17): 1855–1867, 2003. 2, 8

[38] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. 3

[39] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 2020. https://distill.pub/2020/circuits/early-vision. 1, 3

[40] Cynthia Owsley and Michael E Sloane. Contrast sensitivity, acuity, and the perception of'real-world'targets. *British Journal of Ophthalmology*, 71(10):791–796, 1987. 3

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[42] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017. 6

[43] Daniel Porter. Visual acuity, 2022. 3

[44] Mary C Potter. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976. 2, 3, 4

[45] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams. *Neuroscience. 2nd edition.* Sinauer Associates, 2001. 1, 2, 3

[46] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019. 1

[47] Guillaume A Rousselet, Simon J Thorpe, and Michele Fabre-Thorpe. How parallel is visual processing in the ventral pathway? *Trends in cognitive sciences*, 8(8):363–370, 2004. 2, 3, 4

[48] Gary S Rubin and Kathleen Turano. Reading without saccadic eye movements. *Vision research*, 32(5):895–902, 1992. 2, 3, 4

[49] Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, 34:739–751, 2021. 8

[50] 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. 4

[51] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582): 520–522, 1996. 2, 3, 4

[52] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 1

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[54] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022. 3

[55] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022.

[56] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 3

[57] Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. Assaying out-of-distribution generalization in transfer learning, 2022. 5

[58] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 1

[59] Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In

NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future, 2021. 5

[60] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[61] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection, 2022. 5

[62] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Young Joon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. 5

[63] Davide Zoccolan, Minjoon Kouh, Tomaso Poggio, and James J DiCarlo. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27(45):12292–12307, 2007. 2, 3, 4

COLORSENSE: A Comparative Study on Color Vision in Machine Visual Recognition

Supplementary Material

7. Training Details of CIFAR10 Models

We train Resnets/Densenets/MobileNetV2 for 250 epochs with stochastic gradient descent (SGD), weight decay of 0.0005, an initial learning rate of 0.1 and milestone learning rate decay at epochs [150, 200] with a decay factor of 0.1. For ViTs, we follow the procedure in [15] to fine-tune ViTs (preVit) on resized CIFAR10 with SGD, learning rate of 0.003 and no weight decay. We use these training setups and models to evaluate them throughout the paper. Furthermore, as ViT are pre-trained on Imagenet, for fairness, we also include pre-trained Resnets (preResnet) and Densenets (preDensenet) that emulate the protocol in [29] to fine-tune with milestone learning rate decay schedule at one-third and two-thirds of the process. All models are trained with a batch size of 128[1]. In this work, generalization performances are not our foci, so we do not tune any hyperparameters.

We emphasize that we do not compete for test set accuracy in this work, so generalization performance is not our concern as long as it is reasonable. Here we detail our training procedures and their respective performances.

Vision Transformers (preVit224). We use [58] for ViT implementation. [15] fine-tunes models on CIFAR images resized to resolution 384 and a batch size of 512 for 10k steps (about 25 epochs). Due to our computation-resource limit, we instead resize to resolution 224 and use a batch size of 128. Essentially we prolong the training. Using our recipe, for Tiny ViT, we achieve 97.83% on CIFAR10 respectively, which is not far away from ViT-B/16 reported in [15]; for Small ViT, we even achieve accuracy 98.73% on CIFAR10, even better than what is reported in [15]. Note that the model size of Tiny ViT or Small ViT is smaller or equal to what is used in [15].

Pre-trained Resnets (preResnet). We use the pre-trained weights released by PyTorch[41]. The original recipe in [29] is to fine-tune Resnets with a batch size of 512 for 10k steps and learning rate decay at $\frac{1}{3}$ and $\frac{2}{3}$ through the process. To have fair procedures as ViTs, we follow the same setups but not resizing. We use a batch size of 128 to fine-tune on CIFAR for 25 epochs and decay learning rate at 8th and 16th epoch. We found using our recipe is much better than using the original recipe: with the original recipe we achieve accuracy 89.49% on CIFAR10, but our recipe can reach 93.7%.

Pre-trained Densenets/MobilenetV2 (preDensenet/preMobilenetV2) For fair comparison, we follow the procedure as in preResnet and achieve 95.32%, 93.73% respectively.

Resnets/Densenets/MobilenetV2. We achieve accuracy 95.16%/95.85%/92.73% on CIFAR10.

8. Color Pairs

We consult experts in the school of fine arts and use the WAGC ratio as a guide and threshold to define the color discrimination groups. Tab. 4 lists the color pairs and the corresponding difficulty level. If either foreground or background color is labeled as 'others,' we assign 'others' as its color discrimination group as well.

9. Vehicle Classes

The 49 land vehicle classes from ImageNet we considered in this work are: *Model T, ambulance, amphibian, beach wagon, bicycle-built-for-two, bobsled, cab, convertible, crane, dogsled, fire engine, forklift, garbage truck, go-kart, golfcart, grille, half track, harvester, horse cart, jeep, jinrikisha, lawn mower, limousine, minibus, minivan, moped, motor scooter, mountain bike, moving van, oxcart, passenger car, pickup, plow, police van, racer, recreational vehicle, school bus, snowmobile, snowplow, sports car, streetcar, tank, thresher, tow truck, tractor, trailer truck, tricycle, trolleybus, unicycle*.

10. Small Scale Human Study

We record participants' reaction time for each image. Fig. 12 shows that humans also take a slightly longer time to classify objects for harder CD groups, confirming that color vision plays a role in human vision too but in the form of response time. We conclude that humans are more robust to color vision than DNNs, as measured by classification accuracy, and indeed we acknowledge that reaction time can be used as a metric to better understand human's dependence on color vision.

[1] In [29] they use a batch size of 512 but we find that the results are not good, and for a fair comparison, we finally choose a batch size as 128.

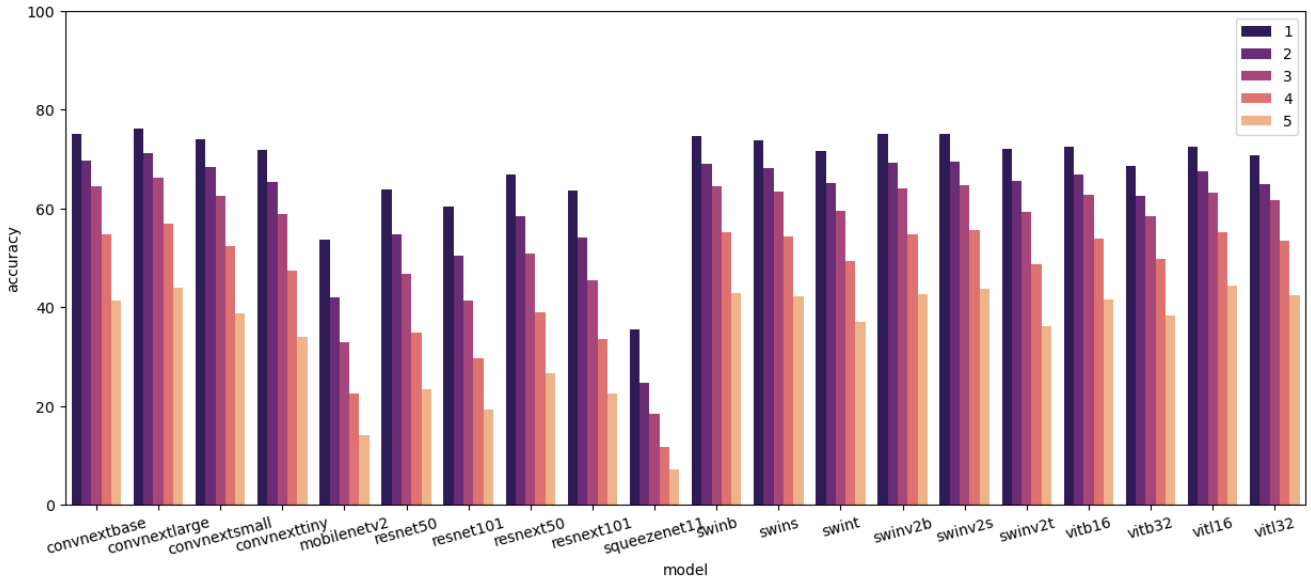
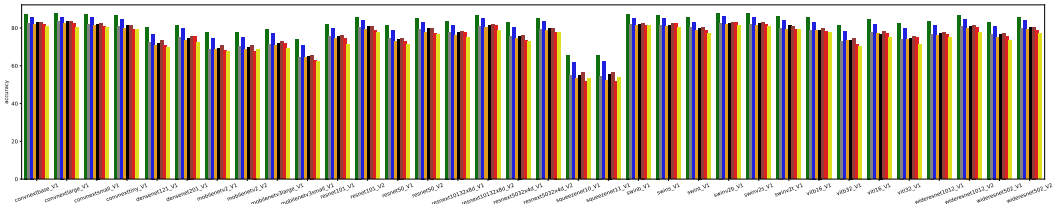
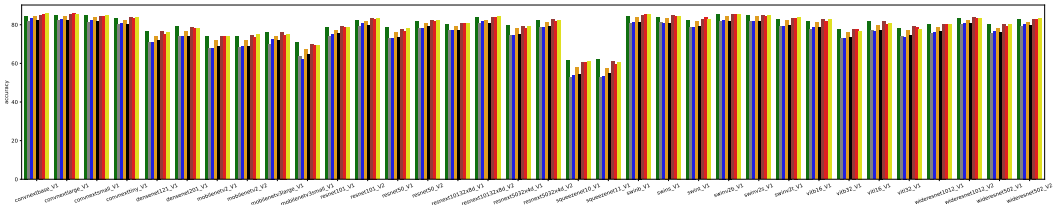


Figure 10. **Performance of each architecture on ImageNet-C by severity.** As an another way to validate our COLORSENSE dataset, the color discrimination groups similarly affect performance as corruption severity (see Fig. 3 for comparison).



(a) Background color as a spurious correlation.



(b) Foreground color as a spurious correlation.

Figure 11. Background/Foreground color as a spurious correlation. The "subgroup degradation" phenomenon appears with all models.

11. Practicality of Proposed sCVR & sCR Metrics

Firstly, we want to highlight the generality of how sCVR was formulated. sCVR is defined to measure the variability of model performance between subgroups while taking accuracy into account. In this work, we have 3 different color discrimination groups per our design; however, in other scenarios, researchers may be interested in a different factor, so they have different subgroups. Therefore, we can design a new metric in a general way, such as sCVR to broadly assess the between-subgroup variability. For instance, we similarly defined an sCR (formula above) for measuring robustness to image corruption (the y-axis in Fig. 9-*left*). We believe that solely using "total accuracy" to evaluate models does not provide sufficient granularity, and a performance variability metric is needed for any study design that involves subgrouping. A simple statistical variance does not suffice due to subgroup sample imbalances. We also designed sCVR along with sCR to account for total accuracy since it is commonly regarded as the primary metric measuring model performance, as the reviewer suggested.

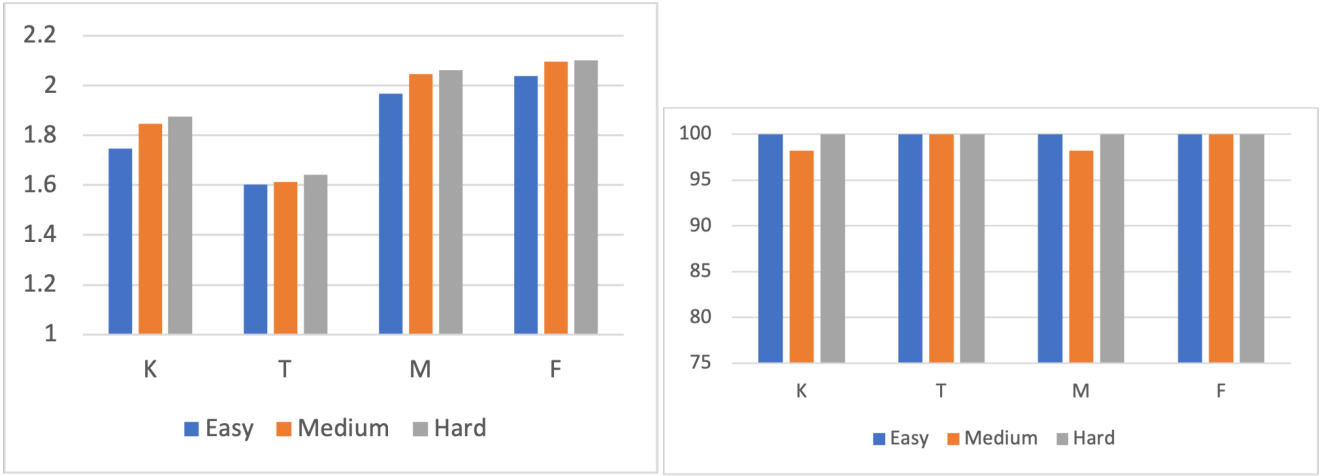


Figure 12. **Human subject experiments.** *Left:* average response time (s) for each color discrimination group on class n02056570, n02128757, n03095699. *Right:* in our experiments, human subjects are not affected by color vision in terms of accuracy when given enough time.

Secondly, to validate the practicality of our metric, we want to highlight that in Fig. 9-*left*, our finding on model robustness to image corruption is consistent with [22] — size-wise, larger models (represented by larger ticks) have higher sCR (more robust). We believe that sCVR and sCR can serve as general metrics for model robustness.

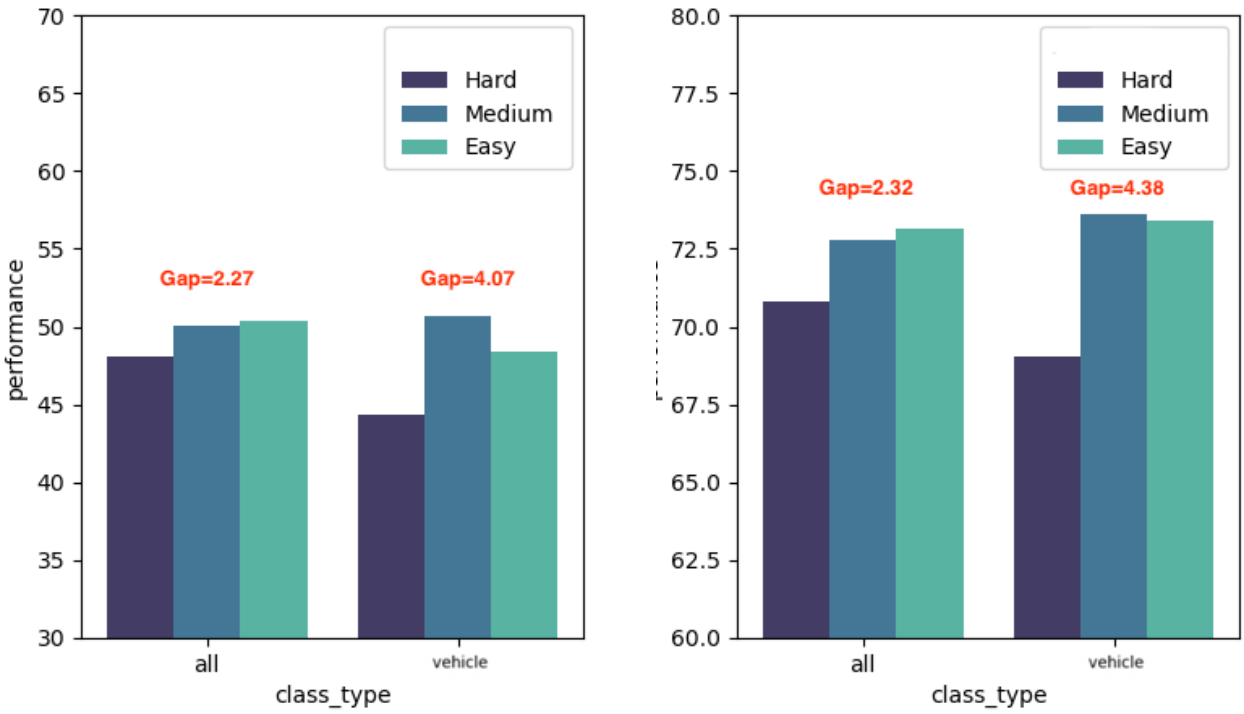
One can also consider model sizes, but that is out of the scope of our work. We will release our dataset to facilitate future studies.

12. Dataset Release

The ColorSense datasets will be released upon the acceptance of this paper.

13. Label Quality

The color annotations are labeled by an individual with a proficient understanding of computer vision to ensure consistency, and they undergo two rounds of verification. As a quality assurance measure, two individuals with substantial technical expertise assessed 500 randomly selected images. The level of consensus reached was 92.4 percent, and after further discussion, an additional 4.8 percent of images initially lacking agreement aligned with the labels we employed for experiments. The rate of disagreement is lower than the typical error rate found in contemporary datasets [38]. We intend to make the dataset publicly available and encourage the community to contribute updates to the color labels.



(a) IoU-70

(b) IoU-50

Figure 13. **Object localization accuracy (%) on all ImageNet classes vs. land vehicle classes.** The performance gap between HARD and EASY groups for vehicle classes are consistently larger than all classes, suggesting vehicle classes are more affected by the color vision effect and therefore possessing safety-related issues. The results are also consistent with the case study in classification task (Sec. 4.3).

Background	Foreground	WAGC ratio	COLORSENSE
green	others	0.0	Others
gray	others	0.0	Others
blue	others	0.0	Others
white	others	0.0	Others
black	others	0.0	Others
brown	others	0.0	Others
red	others	0.0	Others
others	others	0.0	Others
green	green	1.0	Hard
gray	gray	1.0	Hard
blue	blue	1.0	Hard
white	white	1.0	Hard
black	black	1.0	Hard
brown	brown	1.0	Hard
red	red	1.0	Hard
green	gray	1.33	Medium
green	white	1.37	Medium
blue	brown	1.39	Medium
brown	red	1.55	Medium
gray	white	1.82	Medium
blue	red	2.15	Medium
gray	red	2.2	Medium
blue	black	2.44	Medium
green	red	2.91	Medium
black	brown	3.39	Medium
gray	brown	3.41	Medium
white	red	4.0	Medium
green	brown	4.52	Easy
gray	blue	4.72	Easy
black	red	5.25	Easy
white	brown	6.2	Easy
green	blue	6.26	Easy
blue	white	8.59	Easy
gray	black	11.54	Easy
green	black	15.3	Easy
white	black	21.0	Easy

Table 4. **Color pairs.**