ARTICLES
https://doi.org/10.1038/s42256-021-00437-5
nature machine intelligence

When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations

Spandan Madan[1,2] ✉, Timothy Henry[2,3], Jamell Dozier[2,3], Helen Ho[4], Nishchal Bhandari[4], Tomotake Sasaki[5], Frédo Durand[4], Hanspeter Pfister[1] and Xavier Boix[2,3] ✉

Object recognition and viewpoint estimation lie at the heart of visual understanding. Recent studies have suggested that convolutional neural networks (CNNs) fail to generalize to out-of-distribution (OOD) category–viewpoint combinations, that is, combinations not seen during training. Here we investigate when and how such OOD generalization may be possible by evaluating CNNs trained to classify both object category and three-dimensional viewpoint on OOD combinations, and identifying the neural mechanisms that facilitate such OOD generalization. We show that increasing the number of in-distribution combinations (data diversity) substantially improves generalization to OOD combinations, even with the same amount of training data. We compare learning category and viewpoint in separate and shared network architectures, and observe starkly different trends on in-distribution and OOD combinations, that is, while shared networks are helpful in distribution, separate networks significantly outperform shared ones at OOD combinations. Finally, we demonstrate that such OOD generalization is facilitated by the neural mechanism of specialization, that is, the emergence of two types of neuron—neurons selective to category and invariant to viewpoint, and vice versa.

The combination of object recognition and viewpoint estimation is essential for effective visual understanding. In recent years, convolutional neural networks (CNNs) have offered state-of-the-art solutions for both these fundamental tasks[1–8]. However, recent work also suggests that CNNs have a hard time generalizing to combinations of object categories and viewpoints not seen during training: out-of-distribution (OOD) generalization is a challenge. For object recognition, work has shown CNNs struggling to generalize across spatial transformations like two-dimensional (2D) rotation and translation[9–11], and non-canonical three-dimensional (3D) views[12,13]. For viewpoint estimation, previous work has proposed learning category-specific models[5,14] or feeding class predictions as input to the model[15,16], as generalizing to novel categories is a challenging task.

It remains unclear when and how CNNs may generalize to OOD category–viewpoint combinations. Figure 1a presents a motivating example: would a network trained on examples of a Ford Thunderbird seen only from the front, and a Mitsubishi Lancer seen only from the side generalize to predict car model (category) and viewpoint for a Thunderbird shown from the side? If so, what underlying mechanisms enable such OOD generalization?

In this Article, we investigate the impact of two key factors (data diversity and architectural choices) on the capability of generalizing to OOD combinations, and the neural mechanisms that facilitate such generalization. Concretely, we introduce the following discoveries.

The first discovery is that data diversity significantly improves OOD performance, but degrades in-distribution performance. We investigate the role of data diversity by varying the number of in-distribution category–viewpoint combinations, keeping dataset size constant. We find that data diversity matters significantly. For a constant dataset size, increasing data diversity makes the task more challenging, as reflected in the deteriorating in-distribution performance. Yet, increasing data diversity substantially improves performance on OOD combinations.

The second discovery is that Separate architectures significantly outperform Shared ones on OOD combinations unlike in distribution. We also analyse the performance of different architectures in the multi-task setting of simultaneous category and viewpoint classification, that is, learning category and viewpoint in Shared or in Separate (no layers shared) architectures. Our results reveal that Separate architectures generalize substantially better to OOD combinations compared with Shared architectures. Also, this trend is in stark contrast with the trend for in-distribution combinations, where Shared architectures perform marginally better. Thus, the belief that Shared architectures outperform Separate ones when tasks are synergistic should be revisited[17], as their relative performance strongly depends on whether the test sample is in distribution or OOD.

The third discovery is that neural specialization facilitates generalization to OOD combinations. Existing work suggests that OOD generalization is facilitated by selective and invariant representations[18–21]. However, this has not been demonstrated for deep learning, and does not extend to simultaneous category and viewpoint classification. To address this, we propose the neural mechanism of specialization—the emergence of two types of neuron, one driving OOD generalization for category, and the other for viewpoint. This corresponds to neurons selective to a category and invariant

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [2]Center for Brains, Minds and Machines, Cambridge, MA, USA. [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institue of Technology, Cambridge, MA, USA. [5]Fujitsu Laboratories, Kawasaki, Japan. ✉e-mail: spandan_madan@g.harvard.edu; xboix@mit.edu

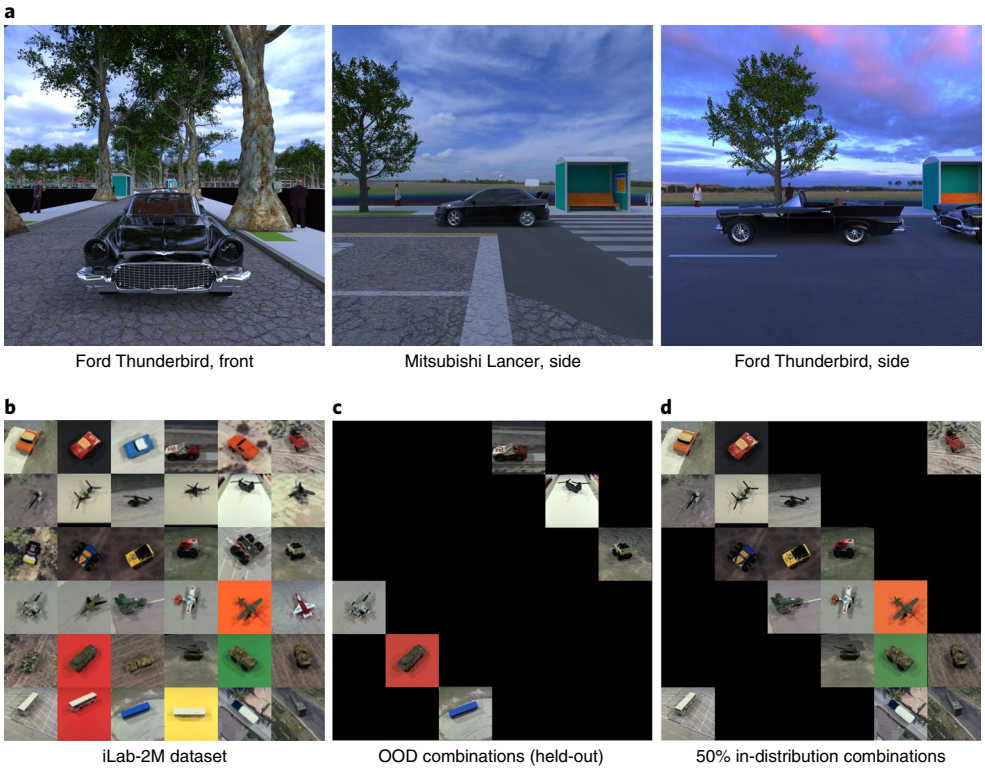

Fig. 1 | Category–viewpoint datasets. a, Our new Biased-Cars dataset: Can a network shown only the Ford Thunderbird from front and the Mitsubishi Lancer from side generalize to classify the category and viewpoint for a Thunderbird seen from the side? **b**, iLab-2M dataset[22,23]. Each cell represents a unique category–viewpoint combination (categories vary between rows, viewpoints between columns) with multiple object instances per category and backgrounds. **c**, Held-out test set of category–viewpoint combinations. Same held-out test set is used to evaluate networks trained with different number of in-distribution combinations. **d**, Biased training set with 50% of category–viewpoint combinations. Number of categories and viewpoints selected is always equal.

to viewpoint, and vice versa. We show that the CNN generalization behaviour trend correlates with the degree of specialization of the neurons.

These results are consistent across multiple CNNs and datasets including the natural image dataset iLab-2M[22,23], variations of MNIST[24,25] extended with position and scale, and a challenging new dataset of car model recognition and viewpoint estimation—the Biased-Cars dataset—which we introduce in this paper. This dataset consists of 15,000 photo-realistic rendered images of several car models at different positions, scales and viewpoints, and under various illumination, background, clutter and occlusion conditions. With this, we hope to provide a first milestone in understanding the underlying mechanisms that enable OOD generalization in multi-task learning for category and viewpoint classification.

Datasets for category–viewpoint classification

Most existing datasets with category and viewpoint labels[13,26–28] present two major challenges: (1) lack of control over the distribution of categories and viewpoints, or (2) small size. Thus, we present our results on the following datasets, which do not suffer from these challenges:

iLab-2M dataset. iLab-2M[22,23] is a large-scale (2 million images), natural-image dataset with 3D variations in viewpoint and multiple object instances for each category (Fig. 1b). The dataset was created by placing toy objects on a turntable and photographing them from six different azimuth viewpoints, each at five different zenith angles (total 30). From the original dataset, we chose a subset of six object categories: bus, car, helicopter, monster truck, plane and tank. In Fig. 1b, each row represents images from one category and each column images from one azimuth angle. All networks are trained to predict one of six category and viewpoint (azimuth) labels each.

MNIST-Position and MNIST-Scale. Inspired by the MNIST-Rotation dataset[29] which adds rotation to MNIST[24,25] images, we created two more variants by adding viewpoint in the form of position or scale. MNIST-Position was created by placing MNIST images into one of nine possible locations in an empty 3-by-3 grid. For MNIST-Scale we resized images to one of nine possible sizes followed by zero-padding. Images of the digit 9 were left out in both these datasets, ensuring nine categories and nine viewpoints classes (total of 81 category-viewpoint combinations). Sample images are available in the Supplementary Section A.1.

Biased-Cars dataset. Building on other multi-view car datasets for viewpoint estimation[30,31], we introduce a challenging new dataset for simultaneous object category and viewpoint classification—the Biased-Cars dataset. Our dataset features photo-realistic outdoor scene data with fine control over scene clutter (trees, street furniture and pedestrians), car colours, object occlusions, diverse backgrounds (building/road textures) and lighting conditions (sky maps). Biased-Cars consists of 15,000 images of five different car models seen from viewpoints varying between 0–90 degrees of azimuth, and 0–50 degrees of zenith across multiple scales. Our dataset offers two main advantages: (1) complete control over the joint distribution of categories, viewpoints and other scene parameters, and (2) unlike most existing synthetic city datasets[27,32,33] we use physically based rendering for greater photo-realism, which has been

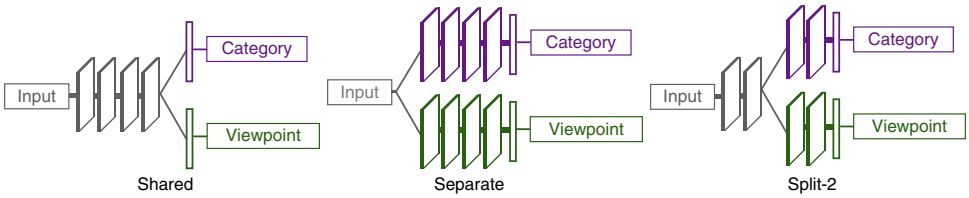

Fig. 2 | Architectures for category recognition and viewpoint estimation. Shared (left), Separate (centre) and Split-2 (right) architectures for ResNet-18. In the Shared architecture, all layers until the last convolutional block are shared between tasks, followed by task-specific fully connected branches. In the Separate architecture, each task is trained in a separate network with no layer sharing. Split-2 presents a middle ground. These architectures are designed similarly for backbones other than ResNet-18.

shown to help networks transfer to natural image data significantly better[34,35]. Sample images are shown in Fig. 1a. As in[26,36], we choose to focus on azimuth prediction. The azimuth is divided into five bins of 18 degrees each, thus ensuring five category (car models) and five viewpoint classes (azimuth bins), for a total of 25 different category–viewpoint combinations. More details can be found in Supplementary Section A.2[37,38].

Additional datasets. In the Supplementary Information, we provide results on two additional standard datasets—MNIST-Rotation[29] and the UIUC3D dataset[39]. Note that the UIUC dataset has a skewed joint distribution of category–viewpoint combinations. This makes it difficult to run controlled experiments. However, the experiments that were possible on this dataset confirm that our findings extend to it as well.

For all datasets, networks are trained to classify both category and viewpoint simultaneously without pretraining, and the number of classes for each task is kept equal to ensure equal treatment. As shown in the experiments, these datasets are challenging benchmarks for testing generalization, with a huge scope for improvement for state-of-the-art CNNs.

Factors affecting generalization behaviour

Below we present the two factors we study for their impact on generalization to OOD category–viewpoint combinations: (1) data diversity and (2) architectural choices.

Generating train/test splits with desired data diversity. All our datasets can be visualized as a square category–viewpoint combinations grid as shown for the iLab dataset in Fig. 1b. Each row represents images from one category, and each column a viewpoint, that is, each cell represents all images from one category–viewpoint combination.

For each dataset, we start by constructing an OOD test split—a set of category–viewpoint combinations are selected and held out from the combinations grid as shown in Fig. 1c. We refer to these as the OOD combinations. Images from OOD combinations are never shown to any network during training. These images are only used to evaluate how networks generalize outside the training distribution. For a fair representation of each category and viewpoint, we ensure that every category and viewpoint class occurs exactly once in the OOD combinations, that is, one cell each per row and column is selected.

Remaining cells in the combinations grid are used to construct multiple training splits with an increasing number of category–viewpoint combinations, that is, data diversity. For each training split, we first sample a set of combinations as shown in Fig. 1d, which we call the in-distribution combinations. Then, we build the training data-split by sampling images from these in-distribution combinations. We ensure that every category and viewpoint occurs equally in the in-distribution combinations, that is, equal number of cells per each row and column. Figure 1d shows the 50% in-distribution training split for the iLab dataset. To ensure that we evaluate the effect of data diversity and not that of data amount, the number of images is kept constant across train splits as the number of in-distribution combinations is increased. Thus, the number of images per combination decreases as the number of in-distribution combinations is increased. Also, note that every network is trained with only one of these training splits at a time, that is, data diversity is kept constant during training.

Architectural choices. One central question addressed in this paper is the impact of architectural choices on the capability to generalize to OOD category–viewpoint combinations. While many separate models have been proposed for object recognition and viewpoint estimation[40,41], recent years have seen a growing trend of architectures inspired by multi-task learning, which suggests that recognition models can benefit from an understanding of object viewpoint, and vice versa[4,5,42–44]. These architectures often learn a shared representation for both tasks, followed by task-specific branches[4,43,45].

Here, we investigate the impact of learning shared representations on the capability of the network to generalize to OOD category–viewpoint combinations, that is, to extrapolate in the multi-task setting of simultaneous category and viewpoint classification. For this, we defined two types of backbone agnostic architecture—the Shared and the Separate architectures. Figure 2 depicts these architectures for a ResNet-18 backbone[1]. In the Shared case, all convolutional blocks are shared between tasks, followed by task-specific fully connected layers, while there are no layers shared between tasks in the Separate architecture. We also investigated three additional Split architectures that represent a gradual transition from Separate to Shared ResNet-18: the Split-1, Split-2 and Split-3 architectures. These were constructed by branching ResNet-18 after 1, 2 and 3 convolutional blocks as shown in Fig. 2. Note that splitting at a layer leads to doubling of the number of neurons in that layer. In our experiments, we show that this increase in width does not provide an advantage.

Generalization through selectivity and invariance

Selectivity and invariance of neurons have long been hypothesized to facilitate generalization in both biological and artificial neural networks[19–21,46–50]. Neurons are commonly interpreted as image feature detectors, such that the neuron's activity is high only when certain features are present in the image[51–55]. We refer to this property as selectivity to an image feature. Selectivity alone, however, is not sufficient to generalize to OOD category–viewpoint combinations. For example, a neuron may be selective to features relevant to a category, but only so for a subset of all the viewpoints. Generalization is facilitated by selective neurons that are also invariant to nuisance features. For instance, in Fig. 1a, neurons that are selective to the Ford Thunderbird and invariant to viewpoint would have very similar activity for the Ford Thunderbird on in-distribution and OOD viewpoints, thus enabling generalization to category recognition.

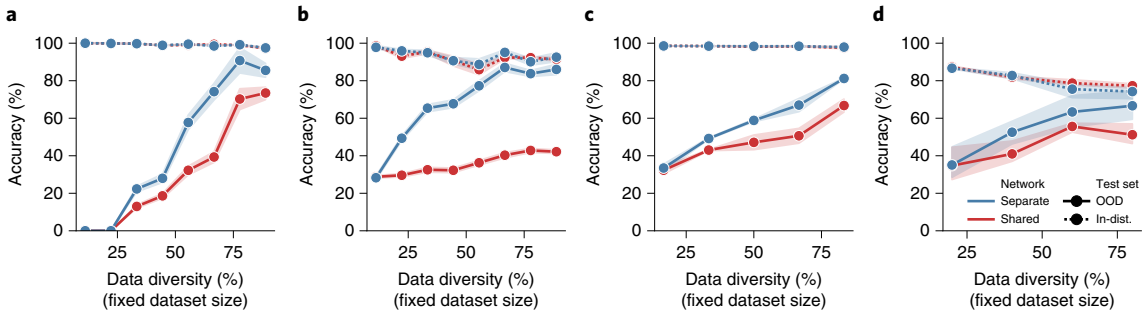

Fig. 3 | Generalization performance for Shared and Separate ResNet-18 as in-distribution combinations are increased for all datasets. The geometric mean of category recognition accuracy and viewpoint estimation accuracy is reported along with confidence intervals (95%). **a**, MNIST-Position dataset. **b**, MNIST-Scale dataset. **c**, iLab dataset. **d**, Biased-Cars dataset.

Similarly, generalization to viewpoint estimation can be enabled by neurons selective to viewpoint and invariant to category.

Here, we present our implementation for quantifying the amount of selectivity and invariance of an individual neuron. Let N be the number of categories or viewpoints in the dataset. We represent the activations for a neuron across all category–viewpoint combinations as an $N \times N$ activations grid, as shown in Fig. 5a. Each cell in this activations grid represents the average activation of a neuron for images from one category–viewpoint combination, with rows and columns representing average activations for all images from a single category (for example, Ford Thunderbird) and a viewpoint (for example, front), respectively. These activations are normalized to lie between 0 and 1 (Supplementary Section B.1). For neuron k, we define a^k_{ij} as the entry in the activations grid for row (category) i and column (viewpoint) j. Below we introduce the evaluation of a neuron's selectivity score with respect to category and invariance score with respect to viewpoint. Viewpoint selectivity score and category invariance score can be derived analogously (Supplementary Section B.2).

Selectivity score. We first identify the category that the neuron is activated for the most on average, that is, the category that has the maximum sum across the rows in Fig. 5a. We call this category the neuron's preferred category, and denote it as $i^{\star k}$, such that $i^{\star k} = \arg\max_i \sum_j a^k_{ij}$. The selectivity score compares the average activity for the preferred category (denoted as $\hat{a}^k$) with the average activity of the remaining categories ($\bar{a}^k$). Let S^k_c be the selectivity score with respect to category, which we define as is usual in the literature (for example, refs. [56,57]) with the following expression:

$$S^k_c = \frac{\hat{a}^k - \bar{a}^k}{\hat{a}^k + \bar{a}^k}, \quad \text{where } \hat{a}^k = \frac{1}{N}\sum_j a^k_{i^{\star k}j}, \ \bar{a}^k = \frac{\sum_{i \neq i^{\star k}} \sum_j a^k_{ij}}{N(N-1)} \quad (1)$$

Observe that S^k_c is a value between 0 and 1, and higher values of S^k_c indicate that the neuron is more active for the preferred category as compared with the rest. Selectivity with respect to viewpoint, denoted as S^k_v, can be derived analogously by swapping indices (i,j).

Invariance score. A neuron's invariance to viewpoint captures the range of its average activity for the preferred category as the viewpoint (nuisance parameter) is changed. Let I^k_v be the invariance score with respect to viewpoint, which we define as the difference between the highest and lowest activity across all viewpoints for the preferred category, that is

$$I^k_v = 1 - \left(\max_j \ a^k_{i^{\star k}j} - \min_j \ a^k_{i^{\star k}j} \right) \quad (2)$$

where the range is subtracted from 1 to have the invariance score equal to 1 when there is maximal invariance. Invariance with respect to category, denoted I^k_c, can be derived analogously.

Specialization score. Generalization to category recognition may be facilitated by neurons selective to category and invariant to viewpoint. Similarly, viewpoint selective and category invariant neurons can help generalize well to viewpoint estimation. This reveals a tension when category and viewpoint are learned together, as a neuron that is selective to category, cannot be invariant to category. The same is true for viewpoint. One way this contradiction may be resolved is the emergence of two types of neuron—category selective and viewpoint invariant, and vice versa. We refer to this as specialization. This hypothesis is well aligned with the findings in ref. [58], which showed the emergence of groups of neurons contributing exclusively to single tasks. Thus, in the context of category recognition and viewpoint estimation, we hypothesize that neurons become selective to either category or viewpoint at later layers as the relevant image features for these tasks are disjoint (the category of an object cannot predict its viewpoint, and vice versa).

To classify neuron k as a category or viewpoint neuron, we compare its selectivity for both category and viewpoint (S^k_c and S^k_v). If S^k_c is greater than S^k_v, then neuron k is a category neuron, otherwise, it is a viewpoint neuron. Since generalization capability relies on both invariance and selectivity, we introduce a new metric for a neuron, the specialization score, I^k, which is the geometric mean of its selectivity and invariance scores, that is

$$I^k = \begin{cases} \sqrt{S^k_c I^k_v} & \text{if } S^k_c > S^k_v \quad (\text{category neuron}) \\ \sqrt{S^k_v I^k_c} & \text{if } S^k_c \leq S^k_v \quad (\text{viewpoint neuron}) \end{cases} \quad (3)$$

Below, we present results that show that the specialization score is highly indicative of a network's performance on OOD combinations.

When do CNNs generalize to OOD combinations?

Below, we summarize our findings from evaluating Separate and Shared architectures when tested on unseen images from in-distribution and OOD category–viewpoint combinations. See Supplementary Section C for experimental details[59,60].

For fixed dataset size data diversity enables better OOD generalization but deteriorates in-distribution performance. Figure 3 presents the geometric mean of category and viewpoint classification accuracy for Separate and Shared architectures with the ResNet-18 backbone, for all datasets. These experiments were repeated three times, and here we present the mean performance with confidence intervals. For fixed dataset size, increasing in-distribution combinations makes the task more challenging as images with each

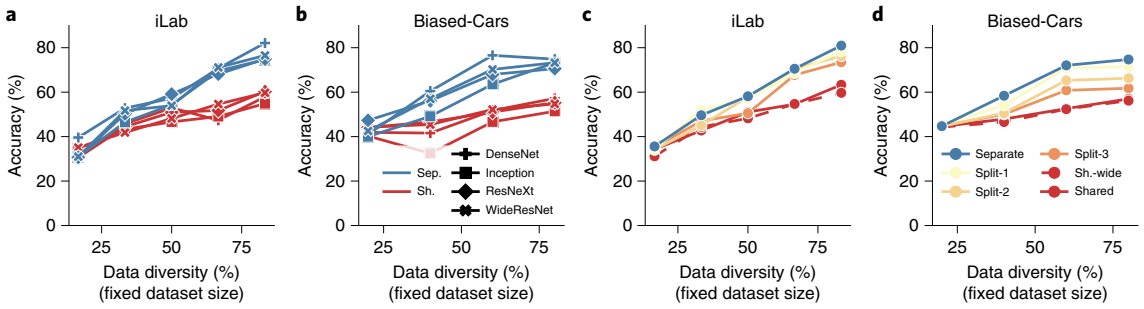

Fig. 4 | Generalization performance for different architectures and backbones as in-distribution combinations are increased for iLab and Biased-Cars datasets. The geometric mean between category recognition accuracy and viewpoint recognition accuracy is reported for OOD combinations as number of in-distribution combinations is increased. **a**,**b**, Accuracy of Separate and Shared for backbones other than ResNet-18, for iLab (**a**) and Biased-Cars (**b**) datasets. **c**,**d**, Accuracy of ResNet-18 Separate, Shared and different Split architectures made by splitting at different blocks of the network, for iLab (**c**) and Biased-Cars (**d**) datasets.

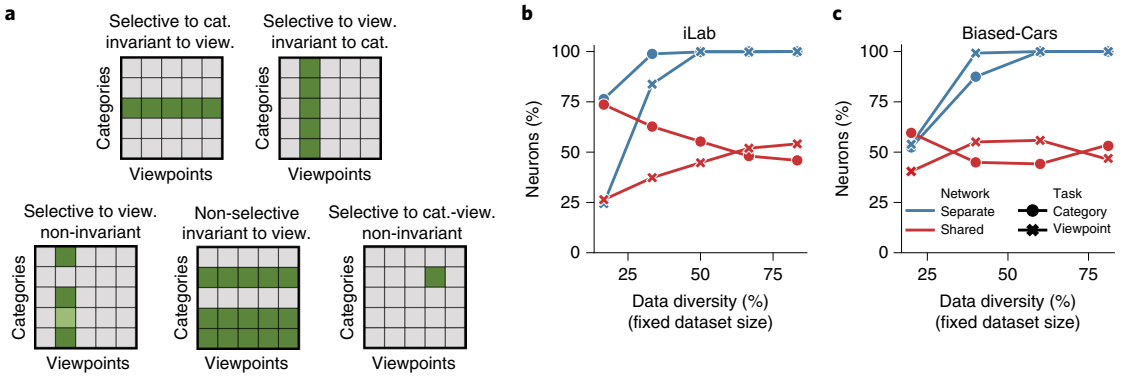

Fig. 5 | Specialization to category recognition and viewpoint estimation. a, Prototypical activation grids for different types of selective and invariant neuron. **b**,**c**, Percentage of neurons after ResNet-18 block-4 that are specialized to category and viewpoint, for iLab (**b**) and Biased-Cars (**c**) datasets. ResNet-18 Separate and Shared networks are evaluated; for Separate, only the task-relevant neurons for each branch are displayed.

category and viewpoint become more diverse, leading to some drop in accuracy on in-distribution combinations. By contrast, both architectures show a significant improvement of their performance on images from OOD combinations, as data diversity increases. We ensured that this result can not be attributed to having closer viewpoint angles between in-distribution and OOD combinations as data diversity is increased (Supplementary Section D.1). CNNs do not theoretically guarantee viewpoint invariance[47], but our result provides reassurance that CNNs can become robust to OOD category–viewpoint combinations as long as they are shown enough diversity during training. Taken together, these results suggest an inherent trade-off between getting better on in-distribution combinations and extrapolating to OOD combinations, which is impacted by training data diversity. Also, these results add to a growing body of work investigating the trade-offs inherent to multi-task learning[61,62].

Even though the geometric mean of category and viewpoint classification increases consistently with increased in-distribution combinations, individual accuracy for these tasks does not always increase consistently (Supplementary Section D.2). We attribute this to the randomness in the selection of in-distribution and OOD combinations. Furthermore, the relative accuracy of the two tasks varies depending on the dataset, and no task is consistently harder than the other across all datasets.

Separate architectures generalize significantly better than Shared ones in OOD combinations unlike in distribution. A striking finding that emerged from our analysis is the contrast in the trends of the in-distribution and OOD performance. While both architectures perform well on new images from in-distribution combinations, Separate architectures outperform Shared ones by a very large margin on OOD combinations. For the ResNet-18 backbone, this result can be seen consistently across all four datasets as shown in Fig. 3. Supplementary Section D.2 shows that Separate also outperforms Shared for category and viewpoint classification individually. Note that previous works have shown that Shared architectures are superior for synergistic tasks, as networks can share features among tasks. These works test on the same combinations as seen during training (in-distribution), and when we do so, we also observe that Shared architectures perform the same or slightly better than Separate ones (Fig. 3 dashed lines). Thus, our results reveal that the relative performance between Shared and Separate depends not only on the synergy between tasks, but also whether the evaluation is in distribution or OOD.

We extended our analysis to Separate and Shared architectures with different backbones, namely ResNeXt[63], WideResNet [64], Inception v3[2] and the DenseNet[3], as shown in Fig. 4a,b. As can be seen, Separate architectures outperform Shared ones by a large margin for all backbones, which confirms that this result is not backbone specific. Investigating further, we experiment with Split architectures, and as can be seen in Fig. 4c,d, there is a consistent, gradual dip in the performance as we move from the Separate to the Shared architectures. Thus, generalization to OOD category–viewpoint combinations is best achieved by learning both tasks separately, with a consistent decrease in generalization as more parameter sharing is enforced.

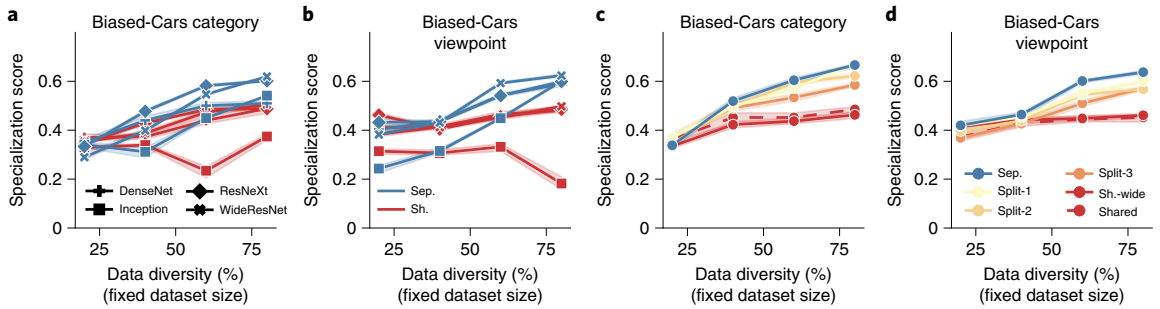

Fig. 6 | Neuron specialization (selectivity to category and invariance to viewpoint, and vice versa) in the Biased-Cars dataset. a,**b**, Median specialization score of neurons (I^k) in Separate and Shared architectures for category (**a**) and viewpoint (**b**) classification tasks, for backbones other than ResNet-18. Confidence intervals (95%) displayed in low opacity. **c**,**d**, Median specialization score of neurons in ResNet-18 Separate and Shared architectures with splits made at different blocks of the network, for category (**c**) and viewpoint (**d**) classification tasks.

To make sure that Separate architectures do not perform better due to the added number of neurons, we made the Shared-Wide architecture by doubling the neurons in each layer of the Shared ResNet-18 network. As Fig. 4c,d shows, this architecture performs very similarly to the Shared one (see additional results in Supplementary Section D.3). This is in accordance with previous results that show that modern CNNs may improve in performance as the width is increased but to a limited extent[65,66].

In the Supplementary Information, we provide a number of additional controls that support the generality of our results. Concretely, we show results for different number of training images (Supplementary Section D.4), viewpoint estimation for four new car models and category prediction for new viewpoints (Supplementary Section D.5), and the order in which category and viewpoint are learned (Supplementary Section D.6). We also present results on additional datasets (Supplementary Section D.7) and architectures (Supplementary Section D.8)[67,68].

How do CNNs generalize to OOD combinations?

We now analyse the role of specialized (that is selective and invariant) neurons in driving generalization to OOD category–viewpoint combinations.

Specialization score correlates with generalization to OOD category–viewpoint. We first investigate the emergence of category and viewpoint neurons in the final convolutional layer of the networks. Figure 5b,c shows the percentage of neurons of each type in Shared and Separate architectures as in-distribution combinations are increased. As can be seen, all neurons in the category and viewpoint branches of the Separate architecture become specialized to category and viewpoint respectively. But in the Shared case, as the network is expected to simultaneously learn both tasks, both kinds of neurons emerge at a ratio of about 50%. We found that this ratio depends on the relative weight of loss terms for the two tasks. When using a different weight from the optimal in terms of maximum geometric mean accuracy, the 50% ratio of specialized neuron becomes unbalanced. For a small number of in-distribution combinations, the ratio of specialized neurons may also be impacted by the relative difficulty of two tasks, with more neurons becoming specialized for the easier task (Supplementary Section E.1).

In Fig. 6 we present the median of specialization scores across neurons, that is, the median of I^k, in the final convolutional layer for Shared, Split and Separate architectures across multiple backbones in Biased-Cars dataset (see Supplementary Section E.2 for results in other datasets). These results are presented separately for the category and viewpoint neurons. We show that as in-distribution combinations increase, there is a steady increase in the specialization score for both category and viewpoint neurons, suggesting specialization. These trends mirror the generalization trends, which suggests that specialization facilitates OOD generalization. Invariance and selectivity scores are reported separately in Supplementary Section E.3. We also show that specialization builds up across layers (Supplementary Section E.4) as expected[20,47].

Separate networks facilitate the emergence of specialized neurons. Figure 6 shows that Separate architectures facilitate specialization, while the Shared architecture makes it harder for the neurons to specialize (lower specialization scores). This might be because unlike the Shared architecture, the branches of the Separate architecture are not forced to preserve features relevant to both tasks. Each branch can develop features that are selective to only one task, and invariant to the other. This may facilitate an increase in specialization and thus enable better performance on OOD combinations. Even though the Shared architecture tries to split into two specialized parts, this specialization is much stronger in the Separate architecture due to already having separate branches.

Conclusions

We have demonstrated that CNNs generalize better to OOD category–viewpoint combinations as the training data diversity grows, for constant dataset size. We have also shown that networks trained separately for category and viewpoint classification surpass by a large margin a shared network trained on both tasks when tested on OOD combinations. We attribute this to the branches in the Separate architecture not being forced to preserve information about both tasks, which facilitates an increase in specialization, that is, selectivity to category and invariance to viewpoint, and vice versa. These results are consistent across five CNN backbones and six datasets, one of them introduced in this paper as a controlled yet photo-realistic benchmark for CNN generalization.

We also found that the aforementioned impact of data diversity and Separate architecture are the opposite for in-distribution and OOD combinations—increased data diversity degrades in-distribution performance, and Separate networks perform worse than Shared ones in in-distribution combinations. This highlights that findings from in-distribution analysis do not apply to OOD.

As a first step towards understanding generalization to OOD combinations, our work makes certain assumptions (summarized in Supplementary Section F), which present interesting directions for future work. These include understanding how generalization is impacted by a larger number of tasks, multiple objects in the image, object symmetries, non-rigid objects and non-uniform ways of holding-out the test set, among others. Finally, we are intrigued to explore what other factors can help learn selective and invariant neural representations which can generalize better and lead the way towards robust, trustable CNNs.

Data availability

To access and cite the Biased-Cars dataset, please visit https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/F1NQ3R&faces-redirect=true.

Code availability

Source code and demos are available on GitHub at https://github.com/Spandan-Madan/generalization_to_OOD_category_viewpoint_combinations.

Received: 11 February 2021; Accepted: 10 December 2021;
Published online: 21 February 2022

References

1. He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
2. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
3. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4700–4708 (IEEE, 2017).
4. Su, H., Qi, C. R., Li, Y. & Guibas, L. J. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proc. IEEE International Conference on Computer Vision* 2686–2694 (IEEE, 2015).
5. Massa, F., Marlet, R. & Aubry, M. Crafting a multi-task CNN for viewpoint estimation. In *Proc. British Machine Vision Conference* 91.1–91.12 (BMVA, 2016).
6. Elhoseiny, M., El-Gaaly, T., Bakry, A. & Elgammal, A. A comparative analysis and study of multiview CNN models for joint object categorization and pose estimation. In *Proc. International Conference on Machine Learning* 888–897 (PMLR, 2016).
7. Mahendran, S., Ali, H. & Vidal, R. Convolutional networks for object category and 3D pose estimation from 2D images. In *Proc. European Conference on Computer Vision Workshops* 698–715 (Springer, 2018).
8. Afifi, A. J., Hellwich, O. & Soomro, T. A. Simultaneous object classification and viewpoint estimation using deep multi-task convolutional neural network. In *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* 177–184 (2018).
9. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L. & Madry, A. Exploring the landscape of spatial robustness. In *Proc. International Conference on Machine Learning* 1802–1811 (PMLR, 2019).
10. Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.* **20**, 1–25 (2019).
11. Srivastava, S., Ben-Yosef, G. & Boix, X. Minimal images in deep neural networks: fragile object recognition in natural images. In *Proc. International Conference on Learning Representations* (2019).
12. Alcorn, M. A. et al. Strike (with) a pose: neural networks are easily fooled by strange poses of familiar objects. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4845–4854 (IEEE, 2019).
13. Barbu, A. et al. ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. *Adv. Neural Inf. Process. Syst.* **32**, 9448–9458 (2019).
14. Tulsiani, S. & Malik, J. Viewpoints and keypoints. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1510–1519 (IEEE, 2015).
15. Xiang, Y., Schmidt, T., Narayanan, V. & Fox, D. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. In *Proc. Robotics: Science and Systems* (2018).
16. Manhardt, F. et al. CPS++: improving class-level 6D pose and shape estimation from monocular images with self-supervised learning. Preprint at https://arxiv.org/abs/2003.05848 (2020).
17. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
18. Giles, C. L. & Maxwell, T. Learning, invariance, and generalization in high-order neural networks. *Appl. Optics* **26**, 4972–4978 (1987).
19. Riesenhuber, M. & Poggio, T. Just one view: Invariances in inferotemporal cell tuning. *Adv. Neural Inf. Process. Syst.* **10**, 215–221 (1998).
20. Goodfellow, I., Lee, H., Le, Q. V., Saxe, A. & Ng, A. Y. Measuring invariances in deep networks. *Adv. Neural Inf. Process. Syst.* **22**, 646–654 (2009).
21. Achille, A. & Soatto, S. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.* **19**, 1947–1980 (2018).
22. Borji, A., Izadi, S. & Itti, L. iLab-20M: a large-scale controlled object dataset to investigate deep learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2221–2230 (IEEE, 2016).
23. Visual variation learning for object recognition. *Jatuporn Toy Leksut* https://bmobear.github.io/projects/viva/ (2016).
24. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
25. *The MNIST Database of Handwritten Digits* (accessed 13 January 2022); http://yann.lecun.com/exdb/mnist/
26. Xiang, Y., Mottaghi, R. & Savarese, S. Beyond pascal: a benchmark for 3D object detection in the wild. In *Proc. IEEE Winter Conference on Applications of Computer Vision* 75–82 (IEEE, 2014).
27. Caesar, H. et al. nuScenes: a multimodal dataset for autonomous driving. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11618–11628 (IEEE, 2020).
28. Min, J., Lee, J., Ponce, J. & Cho, M. Spair-71k: a large-scale benchmark for semantic correspondence. Preprint at https://arxiv.org/abs/1908.10543 (2019).
29. Larochelle, H., Erhan, D., Courville, A., Bergstra, J. & Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proc. 24th International Conference on Machine Learning* 473–480 (PMLR, 2007).
30. Krause, J., Stark, M., Deng, J. & Fei-Fei, L. 3D object representations for fine-grained categorization. In *Proc. 4th International IEEE Workshop on 3D Representation and Recognition* 554–561 (IEEE, 2013).
31. Ozuysal, M., Lepetit, V. & Fua, P. Pose estimation for category specific multiview object localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 778–785 (IEEE, 2009).
32. Qiu, W. & Yuille, A. UnrealCV: connecting computer vision to Unreal Engine. In *Proc. European Conference on Computer Vision* 909–916 (Springer, 2016).
33. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A. & Koltun, V. CARLA: an open urban driving simulator. In *Proc. Annual Conference on Robot Learning* 1–16 (2017).
34. Zhang, Y. et al. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5287–5295 (IEEE, 2017).
35. Halder, S. S., Lalonde, J.-F. & de Charette, R. Physics-based rendering for improving robustness to rain. In *Proc. IEEE/CVF International Conference on Computer Vision* 10203–10212 (IEEE, 2019).
36. Divon, G. & Tal, A. Viewpoint estimation—insights & model. In *Proc. European Conference on Computer Vision* 252–268 (Springer, 2018).
37. Mueller, P. et al. Esri CityEngine—A 3D City Modeling Software for Urban Design, Visual Effects, and VR/AR (Esri R&D Center Zurich, 2020); http://www.esri.com/cityengine
38. Blender—A 3D Modelling and Rendering Package (Blender Foundation, Stichting Blender Foundation, 2020); http://www.blender.org
39. Savarese, S. Fei-Fei, L. 3D generic object categorization, localization and pose estimation. In *2007 IEEE 11th International Conference on Computer Vision* 1–8 (IEEE, 2007).
40. Ghodrati, A., Pedersoli, M. & Tuytelaars, T. Is 2D information enough for viewpoint estimation? In *Proc. British Machine Vision Conference* (BMVA, 2014).
41. Tulsiani, S., Carreira, J. & Malik, J. Pose induction for novel object categories. In *Proc. IEEE International Conference on Computer Vision* 64–72 (IEEE, 2015).
42. Penedones, H., Collobert, R., Fleuret, F. & Grangier, D. *Improving Object Classification Using Pose Information* Technical Report Idiap-RR-30-2012 (Idiap Research Institute, 2012).
43. Zhao, J. & Itti, L. Improved deep learning of object category using pose information. In *Proc. IEEE Winter Conference on Applications of Computer Vision* 550–559 (IEEE, 2017).
44. Li, C., Bai, J. & Hager, G. D. A unified framework for multi-view multi-class object pose estimation. In *Proc. European Conference on Computer Vision* 254–269 (Springer, 2018).
45. Grabner, A., Roth, P. M. & Lepetit, V. 3D pose estimation and 3D model retrieval for objects in the wild. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3022–3031 (IEEE, 2018).
46. Bricolo, E., Poggio, T. & Logothetis, N. K. 3D object recognition: a model of view-tuned neurons. *Adv. Neural Inf. Process. Syst.* **9**, 41–47 (1997).
47. Poggio, T. & Anselmi, F. *Visual Cortex and Deep Networks: Learning Invariant Representations* (MIT Press, 2016).
48. Olshausen, B. A., Anderson, C. H. & Van Essen, D. C. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719 (1993).
49. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
50. Rust, N. C. & DiCarlo, J. J. Selectivity and tolerance (invariance) both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
51. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision* 818–833 (Springer, 2014).

52. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proc. International Conference on Learning Representations Workshop* (2014).
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Object detectors emerge in deep scene CNNs. In *Proc. International Conference on Learning Representations* (2015).
54. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: quantifying interpretability of deep visual representations. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6541–6549 (IEEE, 2017).
55. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 685–694 (IEEE, 2015).
56. Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C. & Botvinick, M. On the importance of single directions for generalization. In *Proc. International Conference on Learning Representations* (2018).
57. Zhou, B., Sun, Y., Bau, D. & Torralba, A. Revisiting the importance of individual units in CNNs via ablation. Preprint at https://arxiv.org/abs/1806.02891 (2018).
58. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
59. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1521–1528 (IEEE, 2011).
60. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
61. Standley, T. et al. Which tasks should be learned together in multi-task learning? In *Proc. International Conference on Machine Learning* (PMLR, 2020).
62. Shin, D., Fowlkes, C. C. & Hoiem, D. Pixels, voxels, and views: a study of shape representations for single view 3D object shape prediction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3061–3069 (IEEE, 2018).
63. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1492–1500 (IEEE, 2017).
64. Zagoruyko, S. & Komodakis, N. Wide residual networks. In *Proc. British Machine Vision Conference* 87.1–87.12 (BMVA, 2016).
65. Nakkiran, P. et al. Deep double descent: where bigger models and more data hurt. In *Proc. International Conference on Learning Representations* (2020).
66. Casper, S. et al. Frivolous units: wider networks are not really that wide. In *Proc. Association for the Advancement of Artificial Intelligence* (2021).
67. Cohen, T. S., Geiger, M., Köhler, J. & Welling, M. Spherical CNNs. In *Proc. International Conference on Learning Representations* (2018).
68. Cohen, T. S., Weiler, M., Kicanaoglu, B. & Welling, M. Gauge equivariant convolutional networks and the Icosahedral CNN. In *Proc. International Conference on Machine Learning* 1321–1330 (PMLR, 2019).

Acknowledgements

We are grateful to T. Poggio and P. Sinha for their insightful advice and warm encouragement. This work has been partially supported by NSF grant IIS-1901030, a Google Faculty Research Award, the Toyota Research Institute, the Center for Brains, Minds and Machines (funded by NSF STC award CCF-1231216), Fujitsu Laboratories (contract no. 40008819) and the MIT-Sensetime Alliance on Artificial Intelligence. We also thank K. Gupta for help with the figures, and P. Sharma for insightful discussions.

Author contributions

S.M., T.H., J.D. and X.B. conceived, designed and implemented the experiments and carried out the analysis, with contributions from T.S., F.D. and H.P.; S.M., H.H., N.B. and F.D. designed and implemented the Biased-Cars dataset; S.M., T.S. and X.B. wrote the manuscript with contributions from F.D. and H.P.; T.S., F.D., H.P. and X.B. supervised the study.

Competing interests

This study received funding from Fujitsu Laboratories. The funder through T.S. was involved in conception of the experiment, writing this article and supervising the study. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00437-5.

Correspondence and requests for materials should be addressed to Spandan Madan or Xavier Boix.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.