scientific reports

Check for updates

OPEN

# Recurrent connections facilitate symmetry perception in deep networks

Shobhita Sundaram[1,4]✉, Darius Sinha[2,4], Matthew Groth[1], Tomotake Sasaki[3] & Xavier Boix[1]✉

**Symmetry is omnipresent in nature and perceived by the visual system of many species, as it facilitates detecting ecologically important classes of objects in our environment. Yet, the neural underpinnings of symmetry perception remain elusive, as they require abstraction of long-range spatial dependencies between image regions and are acquired with limited experience. In this paper, we evaluate Deep Neural Network (DNN) architectures on the task of learning symmetry perception from examples. We demonstrate that feed-forward DNNs that excel at modelling human performance on object recognition tasks, are unable to acquire a general notion of symmetry. This is the case even when the feed-forward DNNs are architected to capture long-range spatial dependencies, such as through 'dilated' convolutions and the 'transformers' design. By contrast, we find that recurrent architectures are capable of learning a general notion of symmetry by breaking down the symmetry's long-range spatial dependencies into a progression of local-range operations. These results suggest that recurrent connections likely play an important role in symmetry perception in artificial systems, and possibly, biological ones too.**

We inhabit a world wherein several entities that carry great ecological significance for us are bilaterally symmetric[1–4]. This includes faces, bodies, animals, and fruits, among many others. The genetic plans of many organisms define symmetric morphologies[5]. The prevalence of symmetric structures in the natural world is complemented by the exquisite sensitivity humans exhibit in detecting such patterns[6,7], a fact that has long been noted by many researchers including Mach[8] and the Gestalt psychologists[9,10]. Humans can discriminate symmetric from non-symmetric patterns even when they are presented tachistoscopically for a fraction of a second, and efficiently search for symmetric patterns amongst non-symmetric distractors[11,12]. Symmetry is an important determinant of the aesthetic rating we assign to a visual pattern, whether that is an abstract structure like the view through a kaleidoscope, or the physiognomy of a person's face[13]. Sensitivity to bilateral symmetry is not unique to humans, but is widespread across the animal kingdom, manifest even in insects and birds[14–17].

The learning of symmetry is particularly interesting in that it represents the acquisition of an abstraction—there is no particular local structure that signals the presence or absence of symmetry, such that two images with very different pixel compositions can both be members of the 'symmetric' class. The criterion that confers class membership is the existence of a relationship between image regions, without imposing any requirements on the contents of the regions themselves. Thus, the classification of a pattern as symmetric requires the assessment of long-range relationships[18–20].

Given the remarkable perceptual sensitivity we have to symmetric patterns, several neuroscientists and computer vision researchers have sought to model the mechanisms underlying this ability[21–28]. However, these models did not investigate the possibility of learning the class of symmetric patterns. This is a significant shortcoming since humans and animals come to exhibit such sensitivity without being given an explicit rule for declaring a pattern symmetric; the ability to acquire symmetric pattern classification through limited experience is a key open avenue for modeling efforts.

The advent of deep networks presents a valuable opportunity in this regard. Convolutional neural networks have exhibited impressive performance on conventional image classification tasks such as object recognition and segmentation[29]. In addition to achieving performance comparable with that of humans in these settings, they are increasingly considered to be reasonable models of object recognition in the human visual system, and to potentially share processing mechanisms with humans[30,31].

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Buckingham Browne & Nichols School, Cambridge, MA, USA. [3]Fujitsu Limited, Kawasaki, Japan. [4]These authors contributed equally: Shobhita Sundaram and Darius Sinha. ✉email: shobhita@mit.edu; xboix@mit.edu

An interesting test of this far-reaching assertion lies in determining whether DNNs are able to learn a rule for detecting symmetry with the same generalization capabilities that humans, or even simpler animals, exhibit. A positive answer would reinforce the claim of DNNs and humans sharing similar representational strategies, while a negative answer would indicate that there may be fundamental differences between the two, notwithstanding the similarities of their performance on conventional classification tasks. More broadly, this exploration will help to determine whether deep networks can learn symmetry, an abstract spatial concept, through exposure to multiple specific exemplars.

Given this state of affairs, we conduct two sets of experiments:

- Experiment Set 1: *Symmetry perception by DNNs modelling brain areas for object recognition.* Since brain areas involved in symmetry perception are shared with brain areas for object recognition[32,33], our first investigation is designed to assess DNNs that have been shown to be effective at modeling the human visual system in terms of object recognition tasks[30,31].
- Experiment Set 2: *Symmetry perception by DNNs with dilated convolutions, recurrence, and transformer networks.* Our second analysis investigates three architecture components that have been designed for problems involving long-range relationships, and are thus especially promising candidates for general symmetry detection. These comprise feed-forward DNNs with dilated convolutions[34], recurrent architectures[35], and transformer networks[36,37].

To evaluate whether the models have learnt a general rule of symmetry, independent of local image features, we test if the DNNs are capable of extrapolating from a limited distribution of training dataset families. We design datasets with varying levels of long-range dependencies and different local image features; we train on a subset of them and then test on the full suite of datasets. Our primary evaluation metric is accuracy on identifying pixel-level symmetry for the out-of-distribution datasets.

We summarize the key differences between the tested architectures across all experiments, and their respective mechanisms for capturing long-range spatial relationships, in Fig. 1. To foreshadow the results, we find that only recurrent networks are able to capture long-range relationships and fully generalize out-of-distribution to novel image families. We also report the real-world applicability of these results by training recurrent networks to recognize symmetry in natural images that include background noise and foreground symmetry. In what follows, we describe both these sets of studies and discuss overall inferences from the compiled results.

## Experiment Set 1: Symmetry perception by DNNs modelling brain areas for object recognition

We assess if feed-forward and recurrent DNNs for object recognition, that are deemed as models of human visual processing, are capable of learning a general rule for symmetry detection. This experiment is founded in previous studies showing that symmetry perception activates in brain areas that are shared with object recognition[32,33]. In particular, we evaluate the following state-of-the-art models: DenseNet, Xception, InceptionResNetV2, InceptionV3, ResNet101, ResNet50, and RCNN-SAT.

**Assessing whether a model has learned a general rule for symmetry.** We design dataset families that are differentiated by the presence and size of an uninformative band of pixels at the center of each image. Images with larger central bands place relevant information at the image edges, thus allowing us to evaluate network recognition of long-range relationships. Examples from each dataset family are visualized in Fig. 2. To assess if networks have learned a general rule for symmetry detection, and are capable of extrapolating to novel instances, we train on a limited distribution of families (band sizes 0 and 4), and test on unseen image families (band sizes {2, 4(dark), 6, 14, 16, 18}); performance on these test families is the metric for assessing generalization. Testing on images with different band sizes ensures that the test images have visual properties that are absent from the training data. All images are constructed as matrices of random noise. Random noise is unbiased to particular shapes, allowing us to evaluate recognition of symmetry with less interference from other visual features.

We examine two training modalities. First, we assess if network representations for object recognition learned from ImageNet already capture symmetry, by transfer-training two fully-connected layers on top of each network with the base layers frozen. We then assess if the networks are capable of learning a generalizable solution by fine-tuning the models end-to-end on our synthetic training families, allowing all layers to update. We find that neither method facilitates learning a general rule for symmetry detection; all trained models struggle to extrapolate to testing families with large band sizes.

**Humans easily detect symmetry in our datasets.** Our first study serves as a simple replication of past studies showing rapid learning of symmetry concept. We test the performance of humans in a symmetry classification task. In the process of doing so, we verify that the concept of symmetry is extractable from our designed stimuli.

For training, ten observers are shown four positive and four negative exemplars with band size 4 (refer to Fig. 2). The positive and negative classes are referred to as 'class 1' and 'class 2', but no explanation is provided as to the class membership criterion, and no mention is made of symmetry. Following this training phase, subjects are shown 50 test images one at a time in five blocks (each block containing five symmetric and five non-symmetric images with band size 4). For each image presented, subjects indicate which class ('1' or '2') it belongs to. The image stays on until the participant has responded. No feedback is provided during the test

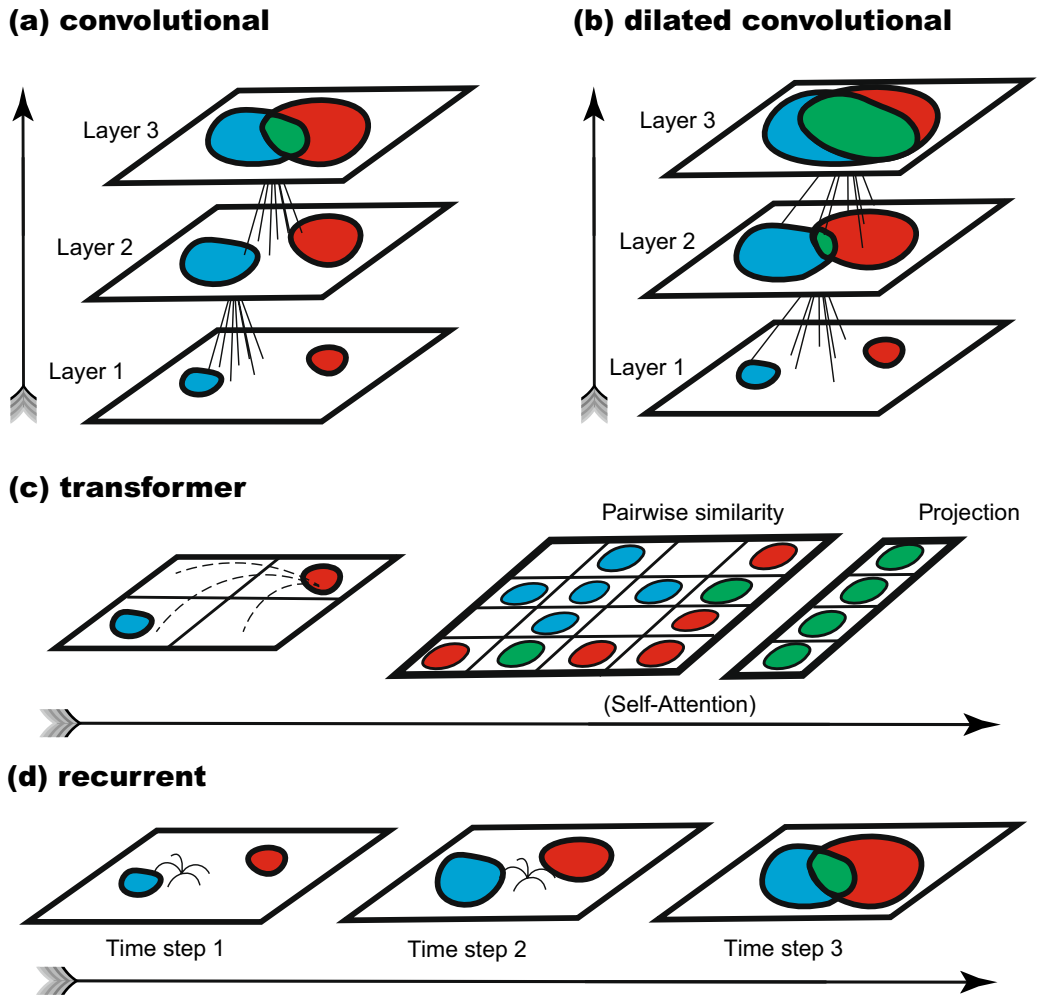



**Figure 1.** Overview of how the architectures evaluated in this work capture long-range spatial dependencies. We indicate in red and in blue the neurons that are influenced by two different, non-local image regions. Also, we indicate in green the neurons that are influenced by both regions and thus could capture dependencies between them. (**a**) The feed-forward convolutional network uses multiple layers to successively expand the receptive field of convolution operations. (**b**) The dilated convolutional network uses atrous convolutions to expand the receptive field with the same number of parameters. (**c**) Transformer architectures use self-attention layers to conduct a pairwise comparison of pixel blocks. (**d**) Recurrent architectures break long-range dependencies into sequences of local operations that are repeated over many time steps (i.e. unrolling steps).

session. After the initial training with two positive and two negative instances, subjects perform near ceiling when classifying the 50 test images. These results indicate that the symmetry property can be extracted with the stimuli that we have created.

**Object recognition DNNs fail to generalize.** We first assess if object recognition representations learned from ImageNet already capture a generalizable notion of symmetry. For each network we freeze the base pre-trained layers and train a classification head for the symmetry detection task. To assess generalization we train on symmetric and non-symmetric images with a limited subset of band sizes {0, 4} and test on images with band sizes {2, 4 (dark), 6, 14, 16, 18} (hyper-parameters are tuned on the validation set of the training families in order to guarantee that the testing families are not used in any way for training, refer to Methods for details). The results are shown in Fig. 3a. We observe that the networks perform poorly across all dataset families, despite showing convergence in training. Performance on novel families is consistently lower than performance on families seen in training. We conclude that the learned representations for object recognition, for both feed-forward and recurrent networks, do not adequately extract features that are relevant to symmetry.

Note that as our DNNs are pre-trained on ImageNet (natural images), the base models may be unable to extract meaningful representations from our synthetic images without further training. Thus, next, we allow the networks to update end-to-end, and examine whether they learn a general rule for symmetry detection from the training distribution.

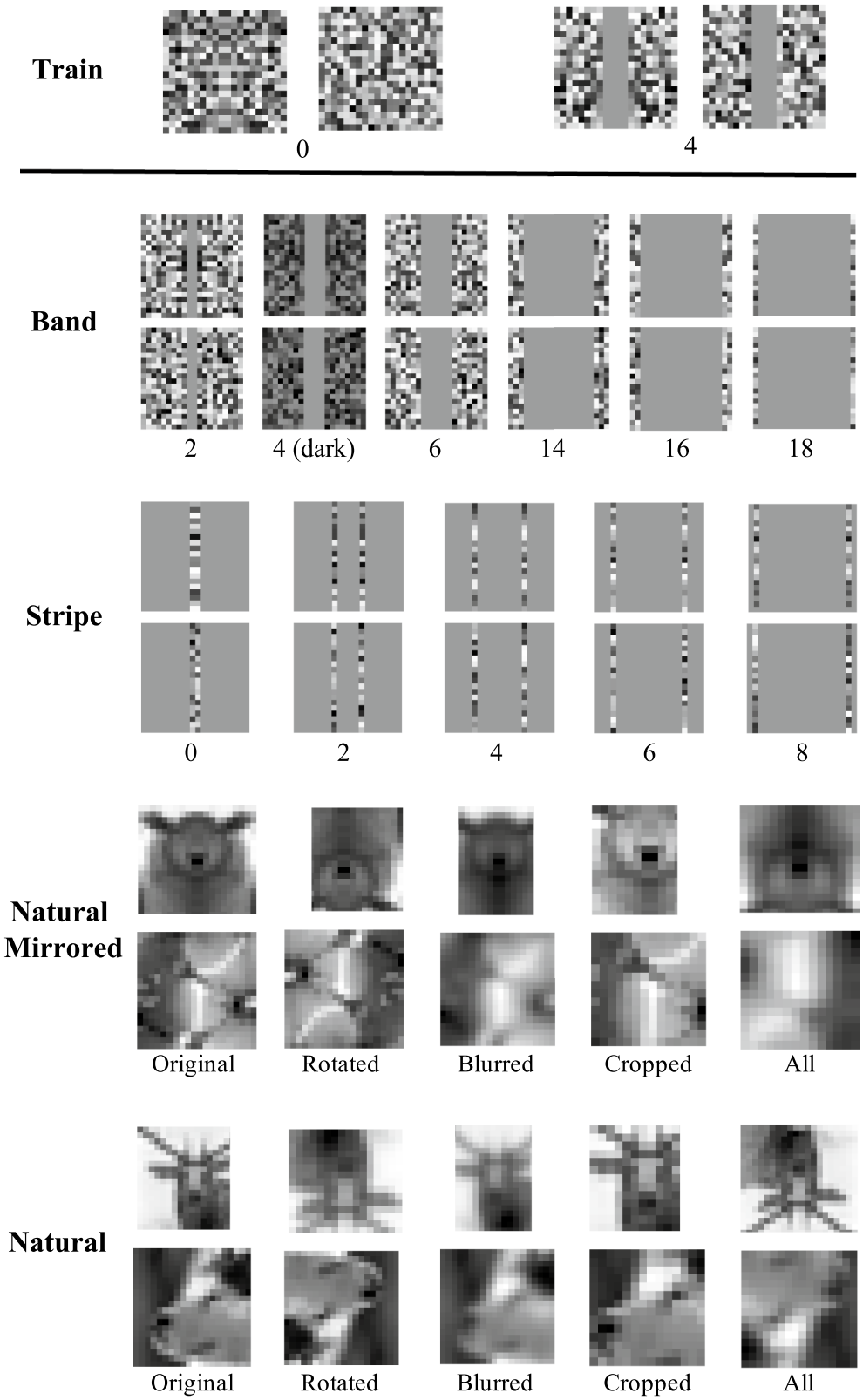



**Figure 2.** Datasets for evaluating learning of a general rule of symmetry detection. The test set images contain visual properties that do not appear in the training set (i.e. expanded central bands, different distances between flanks), thus enabling generalization testing. Note that in Experiment Set 1 we train only with synthetic images; in Experiment Set 2 we experiment with both synthetic and natural training.

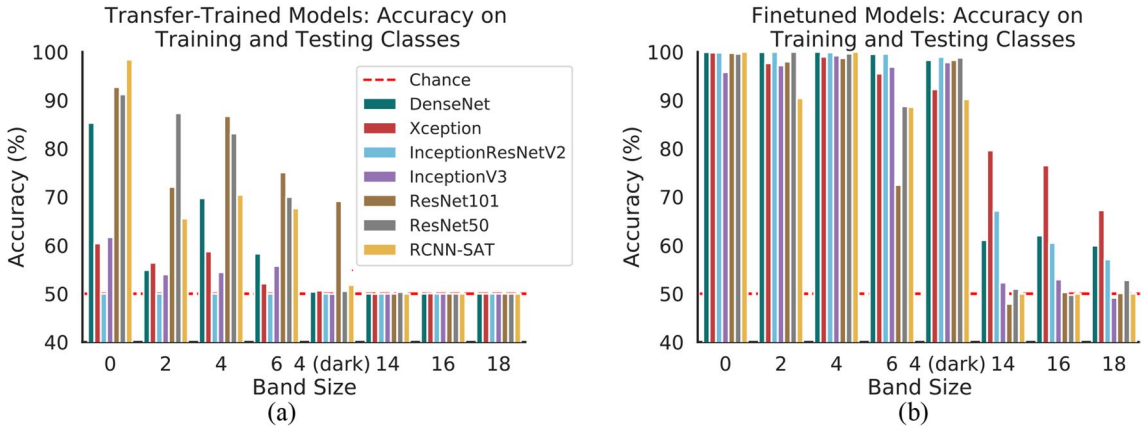



**Figure 3.** Generalization performance of object recognition DNNs. (**a**) Accuracy for six transfer-trained DNNs (with frozen base models) when classifying new exemplars from training classes or instances of slightly different classes. (**b**) Accuracy for six pre-trained DNNs trained with end-to-end fine-tuning when classifying the same exemplars. For both (**b**, **c**) the DNNs do not generalize the categories with large {14, 16, 18}px band sizes.

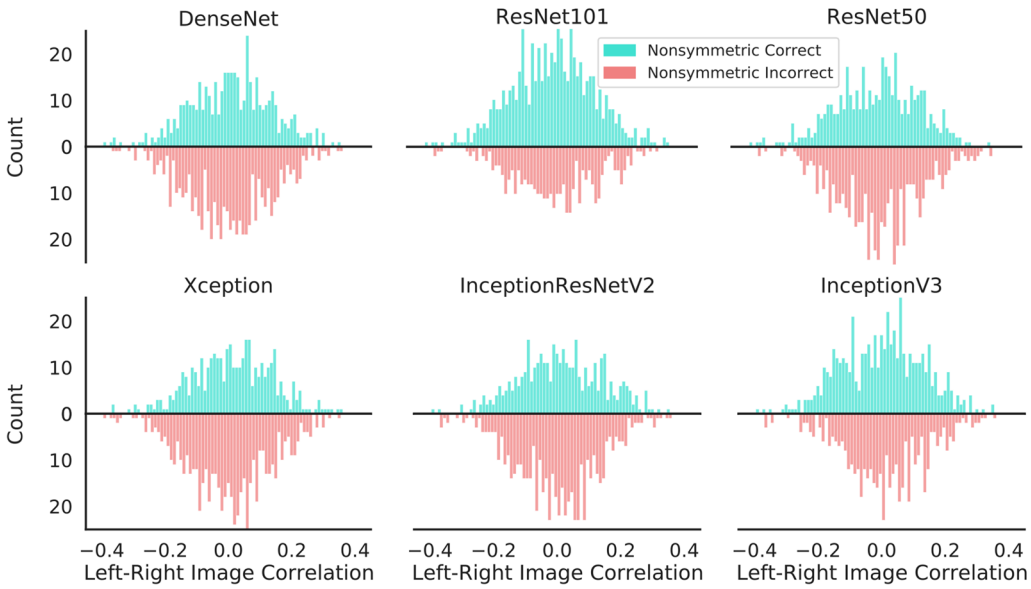



**Figure 4.** Analysis of misclassifications depending on the degree of non-symmetry. We assess the performance of the DNNs on non-symmetric images depending on their degree of non-symmetry. To obtain the degree of non-symmetry, we evaluate the correlation between left and (flipped) right image flanks. We plot the number of images correctly or incorrectly classified depending on such degree of non-symmetry, for images of band size 14. Results indicates that for all DNNs, the probability score assigned to a non-symmetric image is independent of the degree of non-symmetry.

We train and test each network on the same respective datasets; results are shown in Fig. 3b. With the end-to-end fine-tuning the networks demonstrate much stronger performance across the testing families, however are still unable to generalize. Observe that DenseNet, InceptionResNet, and InceptionV3 achieve ceiling accuracy in the training families (band sizes 0 and 4) and are capable of generalizing to some testing families (band sizes {2, 4 (dark), 6}). However all networks perform poorly as the band size expands. For band sizes 14, 16, and 18, for which relevant information is localized to the edges of the images, performance drops by 20 − 50%. We conclude that the networks learn some representation of symmetry, but do not fully capture long-range relationships.

**Analysis of misclassifications.** We next analyze the outputs of the object recognition DNNs to further understand the lack of generalization. In Fig. 4 we examine the accuracy of the networks depending on the degree of non-symmetry of the images. We assess the degree of non-symmetry as the correlation between left and right image flanks (i.e. how close to symmetric an image is). We run this analysis on the fine-tuned networks on non-symmetric images with band size 14, as a representative failure mode of the models. If the networks had

learned a rule for symmetry detection, we might expect that incorrectly classified non-symmetric images would tend to have a lower degree of non-symmetry. However in Fig. 4 we see that this is not the case. For all fine-tuned networks, there is no discernible difference in distribution of the degree of non-symmetry (i.e., left-right correlations) for correctly and incorrectly classified non-symmetric images.

To gain insight regarding the image features learned by fine-tuned DNNs to assess image symmetry, we perform an RSA of the ReLU layer before the network's output[38].

Representations are extracted from 500 symmetric and non-symmetric images for each of the 8 datasets. For every possible pair of images the cosine distance is calculated, i.e., $1 - \frac{1}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}(\mathbf{x}^T \cdot \mathbf{y})$, and then rescaled to [0, 1] to score the dissimilarity between the respective representations. The dissimilarities are displayed in a representational dissimilarity matrix (RDM) in which the images are grouped into the 16 families. Recall that each family contains images that are either symmetric or non-symmetric, with a particular band size and brightness. Thus, the RDM highlights the degree of homogeneity of the representation for each of these features. In Supplemental Fig. 1, we depict the RDM of the different architectures tested in the experiment.

Prototypical models of the RDM are also created to compare our tested DNNs with hypothetical ideal classifiers for each image feature. These prototypical models have assigned dissimilarity scores of either 1 or 0 depending on the feature they classify, as depicted in Fig. 5a. Symmetry-2 is identical to Symmetry-1, except it does not specify the degree of homogeneity when comparing non-symmetric images to other non-symmetric images. This alternative model only needs a strong representation of symmetry, not asymmetry. It may perform just as well in classification tasks. Symmetry-Small-Band is a model which fails with large bands. Thus, the RDM is a function of whether images are symmetric only when comparing images that both have small bands.

We assess the similarity between each of the prototypical models and the neural networks. We do so by gathering Pearson correlation coefficients between the RDMs of each neural network model and each prototypical model. The correlation results can be seen in Fig. 5b. These analyses also include the *LSTM3*, *Dilated*, and *Transformer* models, which are introduced and analyzed in Experiment Set 2. The Symmetry-Small-Band prototype has the highest correlation with the majority of networks investigated in Experiment Set 1. Band-Presence has the highest correlation with InceptionResNetV2 and ResNet101. Thus, the DNNs models tested in this experiment rely on the presence of the band rather than on a general rule of symmetry.

**Discussion.** Our results indicate that state-of-the-art DNNs that have shown impressive performance on object recognition tasks find it challenging to learn a rule for bilateral symmetry detection that can then be applied to images with different visual contents. In our experiments, the networks struggle to extrapolate to images with larger band sizes when trained on images with a subset of possible band sizes. This limitation holds for networks trained both with and without end-to-end fine-tuning. Note that the accuracy drops as the band size increases, indicating that the DNNs do not learn to effectively evaluate long-range relationships. Therefore, it is plausible that our networks have taken shortcuts in the learning process to most efficiently and accurately distinguish between symmetric and non-symmetric images in our training dataset, but failed to extrapolate a general concept of symmetry.

These results are consistent with previous studies demonstrating that object recognition networks struggle to generalize beyond the training distribution, such as object recognition with out-of-distribution object orientations[39,40]. Our study adds to the recent body of works that have suggested that DNNs fail to learn general solutions especially when long-range dependencies and abstract concepts are involved[41].

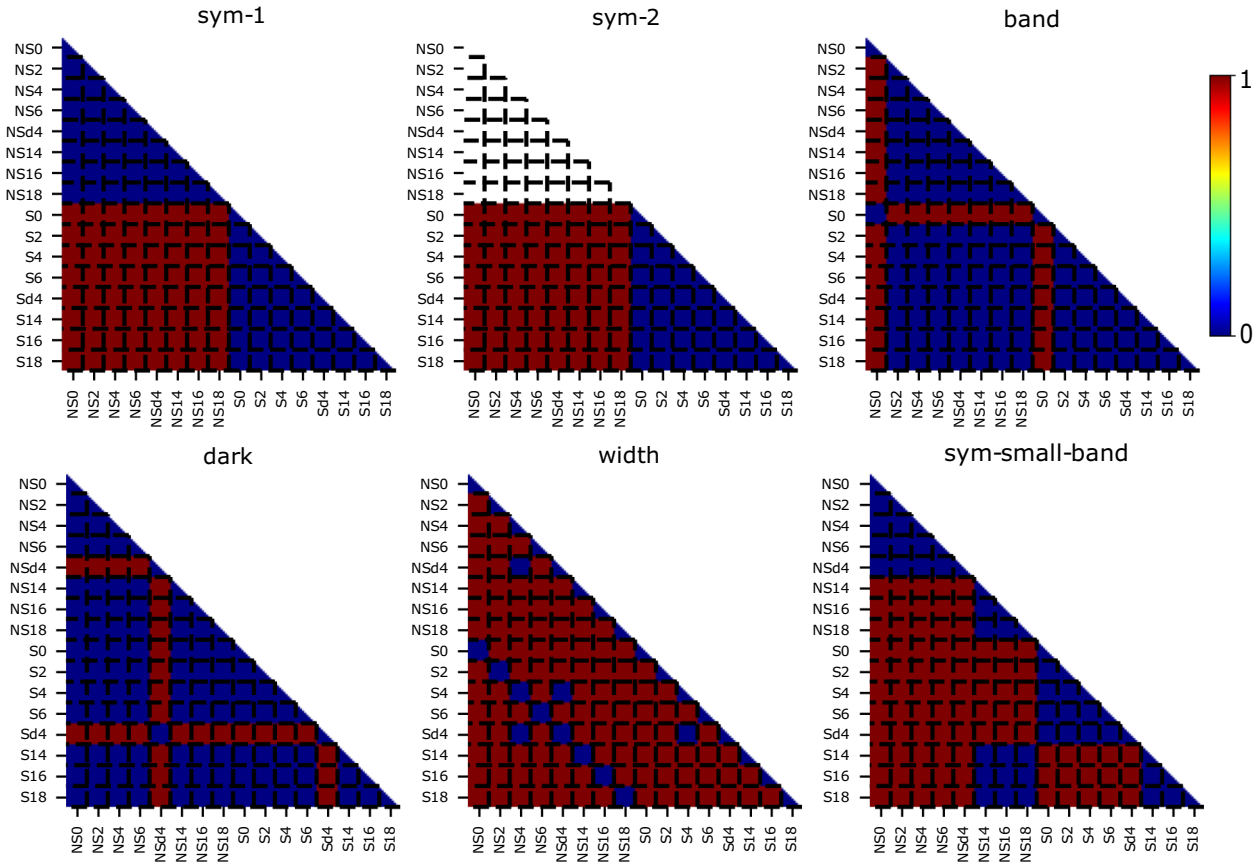
The limited generalization capabilities of the networks may be attributed to the fact that the networks were initially trained for object recognition and as a result, the representations learned may be unable to encode the long-range dependencies required for symmetry perception. Another possible reason is that the network architectures are insufficient to capture long-range dependencies. Note that we evaluate one network with recurrent connections, i.e., RCNN-SAT, as recurrent architectures are better suited to learning representations that capture long-range dependencies than purely feed-forward architectures. RCNN-SAT performs on par with the feed-forward DNNs, suggesting that pre-training for object recognition may be the main cause of the networks' limited generalization capabilities. Yet, the RCNN-SAT recurrent architecture may also not be suitable for symmetry perception.

Our results further call into question the adequacy of these systems as models of human visual perception. It is likely that generalization would improve if images with larger band sizes were included in the training data set. However, even if these object recognition networks were to be able to learn symmetry with much more diverse training sets than the ones we used, that would still not entitle them to be considered good models of biological vision, since animals are able to acquire the concept with very few training instances[14,15,17].

Given the manifest limitation of object recognition networks to acquire the symmetry concept, we shift our focus from networks that are deemed as models of human object recognition to architectures that could allow the networks to gain sensitivity to long-range spatial relationships. The second set of experiments considers three such architectures.

## Experiment Set 2: symmetry perception by DNNs with dilated convolutions, recurrence, and transformer networks

In this second set of experiments, we investigate three architectures that may enable capturing long-range dependencies. These are: the Dilated Convolutional Neural Network (*Dilated*)[34], a three stacked Convolutional LSTM (*LSTM3*)[35], and a Transformer[37]. To our knowledge, none has previously been applied to learning-based symmetry detection. In Fig. 1 and Methods, we describe how these networks are able to accomplish long-range comparisons.




(a)

Pearson Correlation between DNN's RDM and Prototypes' RDM

| | Sym-1 | Sym-2 | Band | Dark | Width | Small-Band |
|---|---|---|---|---|---|---|
| InceptionResNetV2 | 0.13 | 0.11 | **0.34** | -0.14 | 0.13 | 0.33 |
| InceptionV3 | 0.24 | 0.33 | 0.12 | 0.01 | 0.05 | **0.43** |
| ResNet101 | 0.09 | 0.2 | **0.51** | -0.04 | 0.18 | 0.26 |
| Xception | 0.23 | 0.36 | 0.37 | -0.04 | 0.15 | **0.44** |
| DenseNet | 0.25 | 0.39 | 0.38 | -0.02 | 0.2 | **0.54** |
| Dilated | 0.32 | 0.34 | 0.3 | 0.09 | 0.04 | **0.51** |
| Transformer | 0.21 | 0.11 | 0.31 | 0.03 | 0.39 | **0.49** |
| LSTM3 | **0.81** | 0.85 | 0.05 | 0.07 | 0.02 | 0.53 |

(b)

**Figure 5.** Results of representational similarity analysis (RSA). (**a**) Hypothetical patterns showing how the RDMs would look like from perfect classifiers of different features (symmetry, presence of a band, brightness level, or band width). The white area in Symmetry-2 is not included in any correlation calculations. (**b**) Pearson Correlation Coefficients between the RDM of each model (displayed in Supplemental Fig. 1) and the RDM of each prototype.

We analyze the *Dilated* and *LSTM3* architectures directly trained for symmetry perception, without pre-training for object recognition. This facilitates studying the learning of symmetry in isolation, independent of visual cues related to object recognition that may inhibit symmetry perception. Thus we can effectively evaluate the impact of introducing various architectural components. The *Transformer* network could not be trained from scratch in symmetry perception, as it requires hundreds of millions of training examples and an inaccessible amount of computational resources. We use the standard transfer-training procedure for the CLIP model, which is pre-trained using 400 million text-image pairs, leading to the acquisition of powerful general-purpose representations (see Methods). Namely, we freeze the pretrained weights, and train a linear classifier on the symmetry task to evaluate if the pretrained representations capture a general notion of symmetry. In[37] this procedure enables successful zero-shot and few-shot transfer to a plethora of tasks, including those that CLIP was not specifically pretrained for.

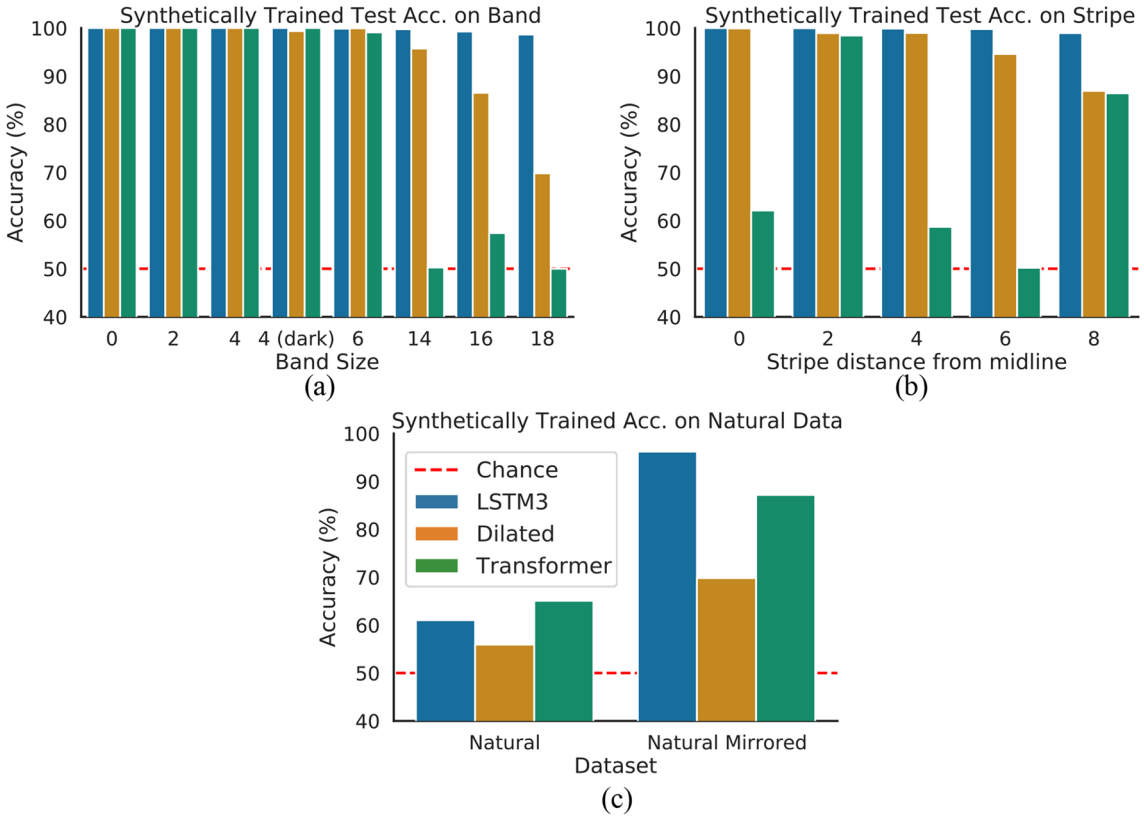



**Figure 6.** Generalization performance of synthetically trained *LSTM3*, *Dilated*, *Transformer*. (**a**) Cross-dataset evaluation accuracies on the Band datasets. (**b**) Accuracies on the Stripe datasets. (**c**) Accuracies on natural image datasets. (**d**) *LSTM3* accuracy on Band datasets for different training data sizes. Across all datasets, only the *LSTM3* fully generalizes.

**Assessing whether a model has learned a general rule for symmetry.** We first assess our networks on the same synthetic datasets as the previously-described object recognition models to compare their respective performances. We then further assess our networks' generalization capabilities by testing them on further novel dataset families (denoted as Stripe datasets) that differ more significantly from the training data. Finally, to assess if the networks are capable of extrapolating to detecting symmetry in the natural world ("in the wild"), we introduce and evaluate on the Natural Mirrored and Natural datasets. For examples from each test dataset refer to Fig. 2.

**Synthetically-trained LSTM3 achieves high generalization accuracy on synthetic test sets.** To demonstrate a systematic improvement in generalization, we first evaluate the performance of the *Dilated*, *LSTM3*, and *Transformer* networks on the Band datasets used in Experiment Set 1. Recall that all networks are trained on symmetric and non-symmetric images with band-sizes 0 and 4, and tested on images with band-sizes {2, 6, 4(dark), 14, 16, 18} to assess generalization accuracy (recall that the hyper-parameters are tuned on the validation set of the training families to guarantee that the testing families are not used in any way for training).

The accuracies across these categories are shown in Fig. 6a. Recall that the object recognition models achieve relatively poor testing accuracies in images with large central bands despite achieving 100% accuracies in training categories, thus demonstrating overfitting. In striking contrast, our *LSTM3* achieves near-perfect accuracy across both training and testing categories. The *Dilated* and *Transformer* networks demonstrate similar trends as the object recognition networks. The *LSTM3* results demonstrate a massive improvement in generalization accuracy (on the same synthetic datasets) compared to Experiment Set 1.

We next evaluate all three networks on test datasets that differ even more greatly from the training distribution. We test on the Stripe family of datasets, for which each image has just a single column of varying pixels on either flank. The Stripe datasets allow us to further evaluate the relationship between network accuracy and the distance between image flanks, thus examining how well the learned solutions evaluate non-local relationships.

The accuracies of the networks for Stripe images are shown in Fig. 6b. Note that across all of these categories, similarly to the Band datasets, the *Dilated* and *Transformer* networks perform worse on categories where the informative image regions are restricted to the edges of the image. This trend is particularly evident for *Dilated*. We conclude that the *Dilated* network's accuracy is dependent on the distance between image flanks, despite its expanded receptive field. While the *Transformer* does not exhibit such a clearly interpretable pattern, it is also clearly does not learn a general rule. In contrast, the *LSTM3* achieves near-perfect accuracy across all categories. The performance of the *LSTM3* is essentially invariant to the distance between image flanks, or the location of

informative regions of the image, providing further evidence that based on the limited distribution of training data it learns a solution that effectively captures long-range relationships.

**Synthetically-trained LSTM3 generalizes to natural mirrored images.** We next test the synthetically trained *LSTM3*, *Dilated*, and *Transformer* on datasets of natural images, denoted as the Natural Mirrored and Natural test datasets (1200 natural images with mirrored left-right flanks and 1200 fully natural images respectively). The difference between these two datasets is that the Natural Mirrored contains symmetric patterns at the pixel level, as there is a one to one correspondence between pixels at corresponding positions of the respective flanks. All datasets previously evaluated in the paper also contain such pixel-level symmetry. In contrast, the Natural dataset evaluates perceived symmetry, as the objects in the image may be symmetric but due to other visual factors (noise, illumination, background, etc.) there may not be a one to one correspondence between pixels at corresponding positions of the respective image flanks.

The results are shown in Fig. 6c. For Natural Mirrored images *LSTM3* achieves near perfect (97%) accuracy while *Dilated* and *Transformer* performs worse, with 70% and 89% accuracy, respectively. These results further strengthen the evidence that the synthetically-trained *LSTM3* captures a solution to detecting pixel-level symmetry, as it even generalizes to distributions of pixels found in the natural world.

All networks perform poorly on the Natural dataset; this result is expected given that the networks were only trained on pixel-level symmetry, while natural images incorporate additional factors that prevents the image flanks from being exactly equal at the pixel-level but that are perceived equal. Later in the paper, we further analyze this phenomenon by training the networks for symmetry detection with natural factors. Next, we analyze the strategies learned by the *LSTM3* network to generalize from a restricted training distribution in pixel-level symmetry detection.

**LSTM3 learns a visual routine for solving symmetry.** We now analyze why the synthetically-trained *LSTM3* is the only network that succeeds in capturing long-range relationships. A theoretical explanation is that *Dilated* solely incorporates an expanded receptive field through atrous convolutions, and such large receptive fields may lead to more complex models that overfit. Meanwhile, the pre-trained representations for visual inputs in the *Transformer* may not capture symmetry, thus preventing transfer-training from learning a general solution. In contrast, *LSTM3* not only expands the receptive field size, but is also capable of breaking long-range dependencies into a sequence of local operations. Recall that the architecture of the *LSTM3* involves applying a feed-forward architecture repeatedly over many time steps steps; the unrolled version of the *LSTM3* is a deep network with shared weights across layers. Thus, with recurrent connections we achieve large-receptive fields while controlling network complexity.

We examine how the distance between image flanks affects the number of time steps required to correctly classify the image as symmetric or non-symmetric. The presence of such a relationship provides crucial insight regarding the mechanism of the *LSTM3*'s learned solution, indicating that it has learned a visual routine composed of local operations that involves expanding outwards from the image center.

To elucidate this relationship, we examine the *LSTM3* testing accuracy across the Stripe categories for different numbers of time steps. Recall that each Stripe image has a single column of varying pixels in either flank that is $x \in \{2, 4, 6, 8, 10\}$ pixels from the midline. In Fig. 7a we show the testing accuracy vs. number of time steps for each Stripe category. We observe that when the image flanks are close together (image stripe 0-2 pixels from the midline) the *LSTM3* requires less than 30 time steps to achieve a high accuracy $>$ 99%. When they are the further apart (image stripe 8 pixels from the midline) 50 timesteps are needed. This result suggests that the *LSTM3* may handle long-range dependencies by propagating information outwards from the center of an image over time.

We gain further insight into the mechanism of the *LSTM3* learned solution by visualizing the neural activations. In particular, for symmetric and non-symmetric images with a band size of 6, we extract the activations of the final *LSTM3* cell. We then perform KMeans clustering on the aggregated activation maps from the same class (with 10 clusters) to elucidate any common patterns. The KMeans clustering algorithm groups the activation maps together into clusters based on each map's Euclidean distance to the nearest cluster centroid. We choose 10 because with this number we observe some redundancy between cluster centroids, indicating that we are seeing a representative range of activity.

The 10 representative activity maps for symmetric and non-symmetric samples respectively are depicted in Fig. 7b. We depict the clusters for activations extracted at 3 different *LSTM3* time steps. Note that the activations taken at the last time step are considered the "output" activations used to decide whether the image is symmetric or not. We observe that at time step 10, the symmetric and non-symmetric activation maps are indistinguishable. At time step 30 the center regions are highlighted more. At the final time step the midline emerges as the primary highlighted region for symmetric centroids, while for non-symmetric centroids the representation patterns propagate uniformly across the whole map. For symmetric samples, by the final time step the centroids are visually homogenous whereas for non-symmetric images the activations do not demonstrate any such pattern. These observations suggest that the identification of the axis of symmetry is important for the *LSTM3*, and support the hypothesis that the network propagates from the center.

These visualizations affirm key differences in how the *LSTM3* represents symmetric and non-symmetric images, and provides further evidence that recurrent connections allow the network to handle long-range dependencies by propagating information over many time steps.

In Experiment Set 1, we introduce the RSA analysis (Fig. 5). We find that most networks except *LSTM3*, have the highest correlation with prototypical classifiers of band presence or symmetry only at small band sizes. In contrast, *LSTM3* has a representation that correlates most with a prototypical symmetry classifier (Symmetry-2).

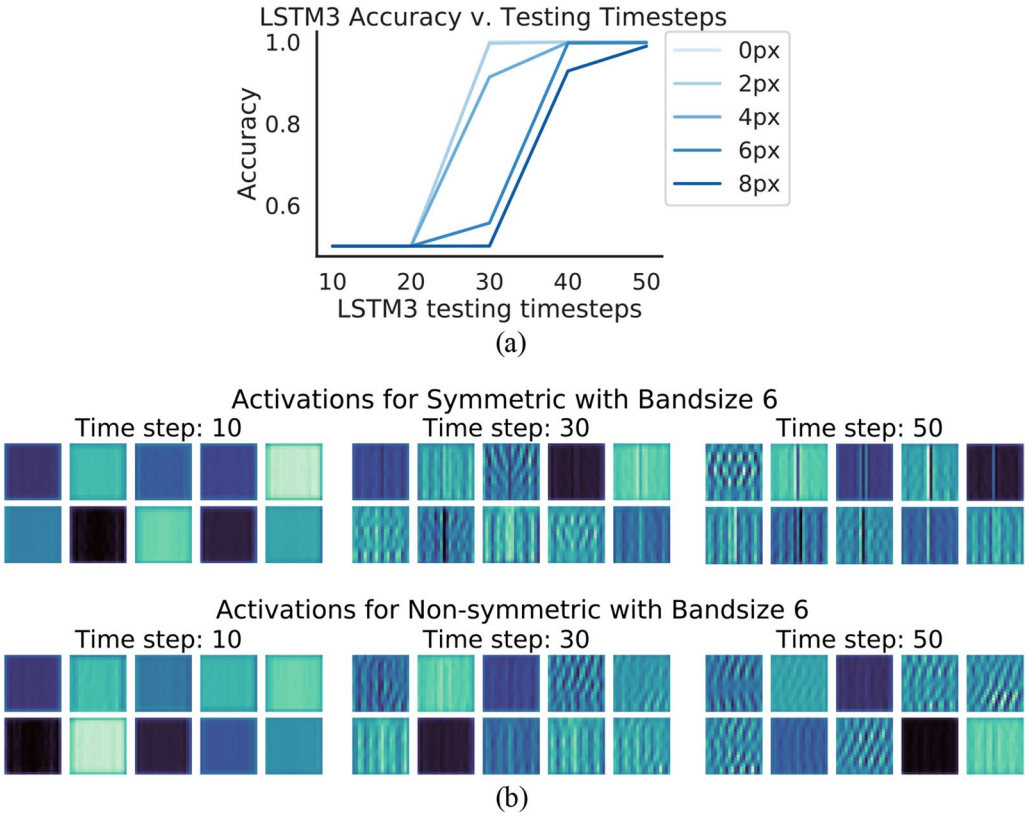



**Figure 7.** Analysis of *LSTM3* representations. (**a**) Plot of the accuracies achieved by the *LSTM3* with restricted numbers of time steps for different categories of Stripe datasets (indicated with the number of pixels between stripes). (**b**) Visualization of 10 representative activation maps from the last *LSTM3* cell at various timesteps, for symmetric and non-symmetric images. Each pixel represents the degree of activation of a neuron in the activation map (darker colors mean values closer to 0 while lighter colors mean larger values). Note that as the timesteps progress, the center axis becomes highlighted for symmetric samples.

**Introducing natural factors inhibits symmetry perception.** Finally, we investigate if the *LSTM3*, *Dilated*, and *Transformer* networks are capable of learning to capturing symmetry in natural images. This task is more complex due to the additional factors (such as noise, background, different illuminations in different parts of the image, etc.) present in natural images that may interfere with the detection of symmetry. Thus, this task requires acquisition of a more general notion of symmetry than a pixel-level one.

We train the *LSTM3*, *Dilated*, and *Transformer* networks on 10800 augmented natural images, using the same learning and hyper-parameter search procedure as synthetic-training experiments (refer to Methods). We then perform cross-dataset evaluation, testing the naturally-trained networks with the highest validation accuracies on all of the test datasets (Band, Stripe, Natural Mirrored, Natural).

In Fig. 8c we show the cross-dataset accuracy for the natural-trained networks on natural images, and observe that the *LSTM3*, *Dilated*, and *Transformer* networks achieve high accuracies (91%, 88%, and 92% respectively for natural data and $>$ 95% for natural mirrored data). In contrast, as seen in Fig. 8a and b, all three networks fail to generalize back to the synthetic test sets. Unlike the results for synthetically-trained networks, the *LSTM3* does not demonstrate a significant performance improvement over either *Dilated* or *Transformer*. Both the *LSTM3* and *Dilated* networks appear to gain some notion of symmetry, performing well above chance accuracy for almost every test dataset, however the full generalization seen with the synthetically-trained *LSTM3* is not evident.

We gain additional insight by examining the accuracy of the network depending on the correlation between left and right (flipped) image flanks (i.e. the degree of non-symmetry of an image) for symmetric natural images that are correctly and incorrectly classified; the results are shown in Fig. 8d. The correctly classified images tend to have a higher left-right correlation (i.e. are more symmetric), while incorrectly classified images have a much wider spread of correlations. These results imply that the networks have picked up some notion of symmetry, and that factors that decrease the visual presence of pixel-level symmetry (evident in decreased left-right correlations) can inhibit the networks' performance.

These results are likely due to the external factors in natural images (noise, illumination, etc) that interfere with the presentation of symmetry, thus making it more difficult to learn symmetry perception from these "in the wild" images. While all networks perform well for test images that are similar to the training set, they struggle with the synthetic data that requires recognition of purely pixel-level symmetry—a visual feature that is more difficult to glean from the natural data.

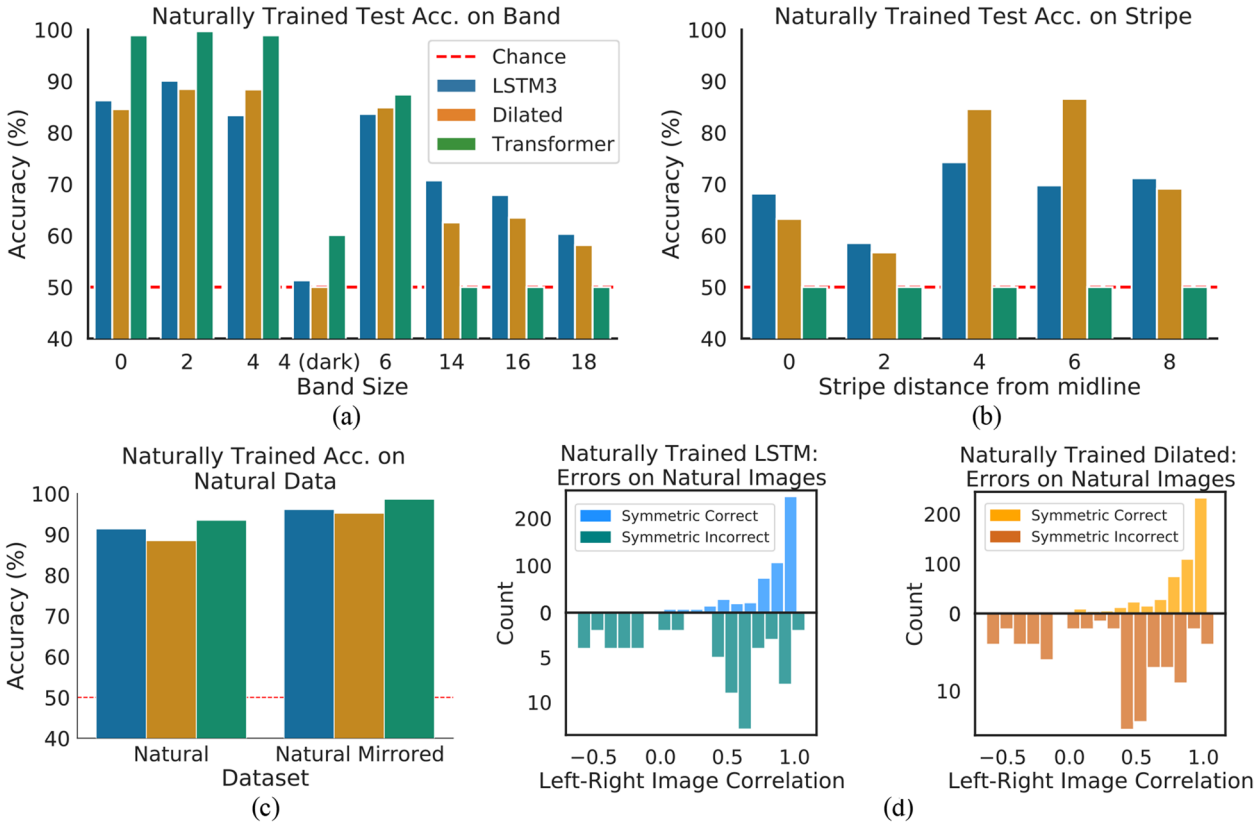



**Figure 8.** Generalization performance of naturally trained *LSTM3*, *Dilated*, *Transformer*. (**a**) Accuracies on the Band datasets. (**b**) Accuracies on the Stripe datasets. (**c**) Accuracies on natural image datasets. (**d**) An analysis of symmetric images correctly/incorrectly classified by *LSTM3* and *Dilated*. The histograms show the number of images correctly or incorrectly classifoed depending on the degree of non-symmetry, which is evaluated as the correlation between the left and right (flipped) halves of the images, where a higher correlation indicates that an image is more symmetric.

## Discussion

Symmetry perception presents two primary challenges. First, it is an abstract feature, indicated by the relationships between pixels, such that a symmetric image with unfamiliar local content should still be recognizable as symmetric. Second, these relationships can be long-range. These challenges are present for both computational and human visual systems; thus symmetry is an excellent assay for studying computational models of human perception.

Our experiments show that only the *LSTM3* model is capable of generalizing to detecting mirror symmetry in novel dataset families. Analysis of the network activations indicate that the *LSTM3* may have learned a routine for symmetry detection that decomposes the long-range dependencies in a sequence of local steps. The *Dilated* and *Transformer* networks, in contrast, struggle with image classes that stress long-range dependencies, and images with different pixel distributions (i.e. natural mirrored images).

These results are consistent with existing works that study the importance of recurrence in modelling long-range dependencies[41–43]. Villalobos et al. have also indicated that recurrent architectures are of critical importance to enable generalization beyond the training distribution in problems with long-range dependencies, in particular for the problem of determining the inside and outside of a closed curve[41]. Our results provide further evidence for the importance of recurrence in enabling DNNs for vision to learn generalizable representations of fundamental visual features that involve long-range relationships. Our results carry implications for downstream tasks, as symmetry in particular is a fundamental visual attribute that has been implicated in facilitating tasks such as pose estimation and depth estimation[44,45].

Furthermore, a body of works in the neuroscience literature argues that recurrence is a crucial component for object recognition[46–50]. Our results demonstrate that recurrent networks enable generalized symmetry perception in a simulated setting. These indicate a possible role for recurrence in human neural processing of long-range dependencies, and in particular symmetry. Such a hypothesis could be experimentally investigated with human EEG readings. If the human brain detects symmetry through a purely feed-forward process, an EEG might take a shorter time to show settled brain activity than if recurrent computations were involved. Our suggested investigation could confirm a functional role of recurrent connections that has not been previously considered, i.e. in handling long-range dependencies given a restricted distribution of training exemplars.

The superior generalization performance observed for the synthetically trained *LSTM3*, however, is not apparent for the *LSTM3* trained on natural images. The natural-trained *LSTM3* successfully generalizes to natural images from a similar distribution, however fails to generalize back to synthetic data, indicating that it has not learned to generally discern symmetry. We also observe that incorrectly classified symmetric natural images tended to have a lower degree of symmetry. These results indicate that while the networks represent some notion of symmetry, natural factors make learning symmetry more difficult. While learning to perceive symmetry in isolation appears to require recurrence -as evidenced by our synthetic-training experiments- learning symmetry detection from "in the wild" images may require other modes of visual processing to account for background asymmetries, noise, different illuminations, and other such factors. Evidently, more investigation of generalizable detection of symmetry in natural images is needed. Future work may involve an architecture search for a model that combines the advantages of recurrence with the advantages of an object recognition model. A promising approach could build upon RCNN-SAT[51], which models to a substantial extent the temporal dynamics of human object recognition with recurrent connections. As suggested by our results, changes in the RCNN-SAT architecture may be required to enable symmetry perception (e.g., the usage of LSTM recurrent connections) and also changes in the training procedure (e.g., training in a richer dataset than ImageNet that requires taking into account long-range dependencies). Additionally, note that this work focuses solely on the detection of bilateral, reflectional symmetry. Further works may investigate other visual phenomena, as well as other forms and axes of symmetry, such as rotational and translational.

In summary, the results presented here compare learned representations of symmetry in object recognition networks, and also DNNs that incorporate mechanisms to tackle long-range dependencies such as dilated convolutions, LSTMs, and Transformer architectures, and evaluate how well these networks learn a generalizable rule for symmetry detection. We concluded that DNNs based on LSTMs facilitate learning a generalized model of symmetry from a limited distribution of training data. Our work motivates future investigations into architectures that incorporate LSTMs to facilitate learning general representations of symmetry from "in the wild" images where symmetry coincides with other visual phenomena.

## Methods

**Experiment Set 1.** *Datasets.* We use specially designed synthetic test datasets that rigorously assess whether trained networks are capable of recognizing pixel-level symmetry—in which the image flanks are perfectly mirrored—for families of images that contain visual properties not apparent in the training dataset. We introduce several families of datasets, split between training and testing families to enable generalization testing.

We train and test on subsets of the Band dataset families. All images are characterized by the size of a central uninformative band. The inclusion of a central band ensures that there are no local features (such as short horizontal segments created by the juxtaposition of identical pixels on the two sides of the axis of symmetry) that can be used to classify patterns as symmetric or non-symmetric.

In particular, for symmetric images with a bandsize $b$, the algorithm fills a $20 \times (10 - \frac{b}{2})$ matrix $L$ with random values drawn from the range $[1, 256]$ to represent the full grayscale spectrum. The matrix is then duplicated and flipped to create matrix $R$. We then create a matrix $B$ of size $20 \times b$, with all values set to 128, that represents the band. The three matrices are concatenated in the order $L + B + R$. For non-symmetric images, we generate separate random matrices of size $20 \times (10 - \frac{b}{2})$ and follow the band-creation and concatenation procedure described above. Each individual image is composed of $20 \times 20$ pixel blocks, where the pixels within each block take the same values. The block size is adjusted for each network to suit the required input size.

Training image families. Our training set is composed of two families of images: one with a central band of size 4, and the other without. In this way, we train on a limited subset of the full distribution (i.e. the full range of band sizes). Each image family is equally split between symmetric/non-symmetric samples. We created a training set with 4000 total images (2000 from each family). We use 90% of the images for training and 10% for validation to tune the hyper-parameters.

Testing image families. Our test set for Experiment Set 1 is composed of six additional image families. The first five feature band sizes $\{2, 6, 14, 16, 18\}$, and the sixth features a band size of 4 with a lower average luminance, achieved by restricting the range of pixel values to $[0, 128]$. Note that each of these families are significantly differentiated from the training families. Different band sizes enable us to evaluate whether the networks are capable of capturing differing sized spatial dependencies, while the last dataset allows us to evaluate if the networks generalize to different luminances.

*Human participants.* The study was approved by the Institutional Review Board at the Massachusetts Institute of Technology (protocol number 0403000050) and all methods were carried out in accordance with relevant guidelines and regulations. All subjects (and their legal representatives, for underage subjects) gave informed consent.

*Networks.* We use 6 feed-forward convolutional DNN architectures, namely DenseNet[52], Xception[53], Inception ResNet V2[54], Inception V3[55], ResNet101, and ResNet50[56]. Each of these networks was originally trained on the ImageNet dataset[57]. These models have been shown to match to a remarkable degree the neural activity and recognition accuracy in primates[31]. We also use RCNN-SAT[51] as it has been shown to mimic human object recognition timing by incorporating recurrent connections, and thus improve previous brain models. We use the pre-trained models that are publicly available.

*Training and hyper-parameter tuning.* We train each network separately for the binary classification problem of symmetry detection. A particular image is classified as 1 if it is symmetric and 0 if it is non-symmetric. For all training procedures we tune the learning rate across {1e−2, 1e−3, 1e−4, 1e−5}. We use a batch size of 32 and train for a minimum of 15 epochs. We perform cross-dataset evaluation using the hyper-parameter set that achieved the highest validation accuracy. In the following, we describe the two different procedures used to train the networks: transfer-training and fine-tuning.

Transfer-training with frozen base models. For all object recognition DNNs we freeze the pre-trained layers and train only the classification head (1 fully-connected 128-node layer) of each network on the training dataset composed of images with band sizes 0 and 4. We then test whether a rule for symmetry is learned using the representations for object recognition. For the RCNN we run the network for eight time steps and average the aggregated outputs, using a single dense classification layer, as described in[51].

End-to-end fine-tuning. We also train end-to-end the networks with the classification head for symmetry detection, such that the weights are initialized as the learned weights from ImageNet, and subsequently all layers are allowed to update based on the synthetic training set. Thus, we follow the same procedure as for transfer-training except that we allow training of all layers in the network.

**Experiment Set 2.** *Datasets.* We introduce more datasets to train the networks in more natural images and also to test the networks in more challenging images. Since we train the networks from scratch, we use an image size of 20 × 20 pixels, which is the smallest possible for our datasets in order to facilitate training as fast as possible.

Training image families. Since we are training the networks from scratch, we use a larger number of training examples than in Experiment Set 1, namely, we use 1e5 training examples. We use the following families to train the networks:

- *Band:* This is the same distributional makeup as the training dataset used in Experiment Set 1. All networks are trained on a dataset composed of symmetric and non-symmetric images with band size 0 and band size 4.
- *Natural:* We additionally conduct experiments for networks trained on image crops from natural images. We combine 176 annotated symmetric images from the NYU Symmetry Database[58] with 75 symmetric images from the CVPR 2013 Symmetry Challenge[59], and 250 non-symmetric images from the ImageNet database[57]. The NYU and CVPR datasets are well-known benchmark datasets for symmetry detection. Using standard data augmentation techniques (cropping, blurring, and rotations) we generate 24 augmented variations of each raw image, leading to a total dataset size of 12000 images. Dividing the dataset with a 90 − 10% split yields a 10800-image training set, and 1200-image testing set. Sample Natural images with data augmentations applied are depicted in Figure 2.

Testing image families. All synthetic datasets contain 1e4 images, and the natural image dataset contains 1200 images.

- *Band:* We test on all of the same testing families introduced in Experiment Set 1 (band sizes {2, 6, 14, 16, 18, 4 (dark)}).
- *Stripe:* In addition, we introduce the Stripe family of datasets to further elucidate how the distance between image flanks impacts perception of pixel-level symmetry, and thus how well the networks capture long-range relationships. A Stripe image contains one column of varying pixels on each flank; the rest of the image pixels are set to the constant value 128 (and are thus uninformative). The dataset subcategories are differentiated by the positions of the two "stripe" columns, as shown in Fig. 2. For a Stripe image with the left-side "stripe" at column $x \in \{2, 4, 6, 8, 10\}$ (where the left-most column is column 0) we fill in a 20 × 20 matrix with value 128. We then replace columns $x$ and $20 - x$ with the same randomly-generated 20 × 1 vector (symmetric images) or different vectors (non-symmetric images). For each possible position of $x$ we generate both a symmetric and non-symmetric dataset.
- *Natural Images:* We additionally test on the test set of natural crops (1200 images).
- *Natural Mirrored dataset:* We create a dataset of images that are derived from the Natural dataset, but are mirrored to have identical left-right flanks. For each image in the 1200 Natural test dataset described above, the right half was replaced by the mirrored left half of that image. Thus these images have the pixel distribution of natural images, but are precisely pixel-level symmetric (unlike fully natural images which symmetry is perceived even though there are variations between the flanks). Sample Natural Mirrored images are shown in Fig. 2.

*Networks.* We explore the following architectures that incorporate mechanisms to tackle long-range dependencies, which are depicted in Fig. 1.

Dilated convolutional neural networks. It is an architecture that employs convolutions with "expanded", i.e., upsampled, kernels[34]. The dilation rate parameter $l$ indicates how much the kernel is expanded, introducing $l-1$ spaces between kernel elements. By increasing the dilation rate monotonically throughout the network layers, the receptive field of the network is expanded while maintaining the same number of parameters, thus facilitating the learning of long-range relationships. We use $3 \times 3$ kernels, 7 layers, and a dilation rate of 4, following the parameters used in[41] as we also found that this architecture works best for symmetry detection.

Transformers. The Transformer model uses a self-attention mechanism to learn long-range dependencies between inputs. Namely, the self-attention mechanisms take as input a set of image patches represented in an embedded space and compare them in a pairwise manner, such that relations across all pairs of image patches can be taken into account, independently on the distance between them. Then, the pairwise similarities between patches are projected into a set of vectors. The transformers stack many layers with self-attention and also standard fully connected layers possibly with skip connections. We use OpenAI's CLIP model, pretrained with 400 million text/image pairs to predict text labels for image inputs[37].

Three stacked Convolutional LSTM[35]. LSTM is a recurrent network that alleviates the well-known issues of training recurrent networks with a large number of unrolling steps. Since it is convolutional, it is suitable for vision problems. A recurrent network can be thought of as a feed-forward network applied repeatedly over many time steps, with shared weights between time steps. For each time step the same image is fed as an input, and the hidden state from the previous time step is carried forward. In the "unrolled" version of the network, time steps are applied as subsequent layers. Here, we stack three convolutional LSTM cells (of 64 channels each) to better facilitate learning a multi-step visual routine for symmetry detection (this was found through an initial pilot experiment in which we assessed architectures with one, two or three cells). In general, recurrent networks are capable of capturing long-range dependencies by breaking them up into sequences of local operations that are repeated over time. Previous research has demonstrated that for some visual problems involving long-range dependencies, a stacked Convolutional LSTM with several cells is capable of learning a simple visual routine that is generalizable to images outside the training distribution[41].

*Training and hyper-parameter tuning.* We train *Dilated* and *LSTM3* from scratch and transfer-train the *Transformer* from the pre-trained architecture on 400 million text/image pairs (we could not train it from scratch on our symmetry datasets, possibly because we did not have enough computational resources to train it with a large number of training examples). Separate experiments are conducted for training all networks on synthetic and natural image sets (as described above). For all experiments we use a 95%/5% split for training/validation. For both synthetic and natural training, we perform cross-dataset evaluation using the network and hyper-parameter set that achieved the highest validation accuracy.

For the *Dilated* and *LSTM3* networks, we test the following hyper-parameters. The convolutional layers use zero-padding, the batch-size is 32, and we explored learning rates $\{1e{-}1, 1e{-}2, 1e{-}3, 1e{-}4, 1e{-}5\}$. For natural training, each network is trained on a 10,800 dataset of natural images, using the training/validation split. For the *LSTM3* we try different numbers of unrolling steps: {5, 10, 20, 30, 50}.

For the *Transformer* architecture, we transfer-train using the standard procedure[37]. Namely, we train a logistic regressor from the representations of the layer before the output using the default parameters (1000 iterations, and regularizer parameter $C = 0.316$, with the Transformer weights frozen).

## Data availability

Our datasets are available at https://dataverse.harvard.edu/dataverse/symmetry. A demo of experiment 1 can be found here: https://tinyurl.com/symmetrydemo.

## Code availability

The code to reproduce the experiments is available at https://github.com/ssundaram21/symmetry.

Received: 5 April 2022; Accepted: 28 November 2022
Published online: 03 December 2022

## References

1. Martindale, M. Q. & Henry, J. Q. The development of radial and biradial symmetry: The evolution of bilaterality. *Am. Zool.* **38**(4), 672–684. https://doi.org/10.1093/icb/38.4.672 (2015).
2. Ball, P. *Shapes* (Oxford University Press, 2009).
3. Manuel, M. Early evolution of symmetry and polarity in metazoan body plans. *C. R. Biol.* **332**(2), 184–209. https://doi.org/10.1016/j.crvi.2008.07.009 (2009).
4. Rosen, J. Symmetry at the foundation of science and nature. *Symmetry.* **1**(1), 3–9. https://doi.org/10.3390/sym1010003 (2009).
5. Davidson, E. & Erwin, D. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800. https://doi.org/10.1126/science.1113832 (2006).
6. Pornstein, M. H. & Krinsky, S. J. Perception of symmetry in infancy: The salience of vertical symmetry and the perception of pattern wholes. *J. Exp. Child Psychol.* **39**(1), 1–19. https://doi.org/10.1016/0022-0965(85)90026-8 (1985).
7. Wenderoth, P. The salience of vertical symmetry. *Perception.* **23**(2), 221–236. https://doi.org/10.1068/p230221 (1994).
8. Mach E. E. *Beiträge zur Analyse der Empfindungen (Contributions to the Analysis of Sensations)* (1886).
9. Wertheimer, M. Untersuchungen zur Lehre der Gestalt. II.. *Psychol. Forsch.* **4**, 301–350 (1923).
10. Koffka, K. *Principles of Gestalt Psychology* (Harcourt, Brace, 1935).
11. Wolfe, J. M. & Friedman-Hill, S. R. On the role of symmetry in visual search. *Psychol. Sci.* **3**(3), 194–198. https://doi.org/10.1111/j.1467-9280.1992.tb00026.x (1992).

12. Norcia, A. M., Candy, T. R., Pettet, M. W., Vildavski, V. Y. & Tyler, C. W. Temporal dynamics of the human response to symmetry. *J. Vis.* **2**(2), 1–1. https://doi.org/10.1167/2.2.1 (2002).
13. Scheib, J. E., Gangestad, S. W. & Thornhill, R. Facial attractiveness, symmetry and cues of good genes. *Proc. Biol. Sci.* **266**(1431), 1913–1917 (1999).
14. Delius, J. D. & Nowak, B. I. Visual symmetry recognition by pigeons. *Psychol. Res.* **44**, 199–212 (1982).
15. Giurfa, M., Eichmann, B. & Menzel, R. Symmetry perception in an insect. *Nature* **382**, 458–461 (1996).
16. Moller, A. & Thornhill, R. Bilateral symmetry and sexual selection: A meta analysis. *Am. Nat.* **151**, 174–92. https://doi.org/10.1086/286110 (1998).
17. Benard, J., Stach, S. & Giurfa, M. Categorization of visual stimuli in the honeybee *Apis mellifera*. *Anim. Cogn.* **9**, 257–70. https://doi.org/10.1007/s10071-006-0032-9 (2006).
18. Dakin, S. C. & Hess, R. F. The spatial mechanisms mediating symmetry perception. *Vis. Res.* **37**(20), 2915–2930. https://doi.org/10.1016/S0042-6989(97)00031-X (1997).
19. Dakin, S. C. & Herbert, A. M. The spatial region of integration for visual symmetry detection. *Proc. R. Soc. Lond. B* **265**(1397), 659–664. https://doi.org/10.1098/rspb.1998.0344 (1998).
20. Saarinen, J. & Levi, D. Perception or mirror symmetry reveals long-range interactions between orientation-selective cortical filters. *Neuroreport.* **11**, 2133–8. https://doi.org/10.1097/00001756-200007140-00015 (2000).
21. Cham, T. & Cipolla, R. *Skewed Symmetry Detection Through Local Skewed Symmetries*. (BMVC, 1994).
22. Stevens, C., Joung, W. & Latimer, C. Modelling symmetry detection with back-propagation networks. *Spat. Vis.* **8**(4), 415–431. https://doi.org/10.1163/156856894X00080 (1994).
23. Tyler, C. W. *Human Symmetry Perception and its Computational Analysis* (Zeist, 1996).
24. Wagemans, J. Characteristics and models of human symmetry detection. *Trends Cogn. Sci.* **1**(9), 346–352. https://doi.org/10.1016/S1364-6613(97)01105-4 (1997).
25. Fukushima, K. & Kikuchi, M. Symmetry axis extraction by a neural network. *Neurocomputing.* **69**, 1827–1836. https://doi.org/10.1016/j.neucom.2005.11.010 (2006).
26. Poirier, F. J. A. M. & Wilson, H. R. A biologically plausible model of human shape symmetry perception. *J. Vis.* **10**(1), 1–16. https://doi.org/10.1167/10.1.9 (2010).
27. Funk, C. & Liu, Y. Beyond planar symmetry: Modeling human perception of reflection and rotation symmetries in the wild. *IEEE Int. Conf. Comput. Vis. (ICCV).* **2017**, 793–803 (2017).
28. George, J. K., Soci, C., Miscuglio, M. & Sorger, V. J. Symmetry perception with spiking neural networks. *Sci. Rep.* **11**(1), 1–14 (2021).
29. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**(7553), 436–444 (2015).
30. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**(11), e1003915 (2014).
31. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**(3), 356–365 (2016).
32. Sasaki, Y., Vanduffel, W., Knutsen, T., Tyler, C. & Tootell, R. Symmetry activates extrastriate visual cortex in human and nonhuman primates. *Proc. Natl. Acad. Sci.* **102**(8), 3159–3163. https://doi.org/10.1073/pnas.0500319102 (2005).
33. Keefe, B. D. *et al.* Emergence of symmetry selectivity in the visual areas of the human brain: fMRI responses to symmetry presented in both frontoparallel and slanted planes. *Hum. Brain Mapp.* **39**, 3813–3826 (2018).
34. Yu, F. & Koltun, V. *Multi-Scale Context Aggregation by Dilated Convolutions. CoRR*. http://arxiv.org/abs/1511.07122 (2016).
35. Shi, X. *et al.* Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* Vol. 28 (eds Cortes, C. *et al.*) (Curran Associates Inc, 2015).
36. Vaswani, A. *et al. Attention is All you Need*. http://arxiv.org/abs/1706.03762.
37. Radford, A. *et al. Learning Transferable Visual Models From Natural Language Supervision*. (ICML, 2021).
38. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
39. Alcorn, M. A. *et al.* Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. *IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR).* **2019**, 4840–4849 (2019).
40. Madan, S. *et al.* When and how convolutional neural networks generalize to out-of-distribution category-viewpoint combinations. *Nat. Mach. Intell.* (2022).
41. Villalobos, K. *et al.* Do neural networks for segmentation understand insideness?. *Neural Comput.* **33**(9), 2511–2549 (2021).
42. Linsley, D., Ashok, A., Govindarajan, L., Liu, R. & Serre, T. Stable and expressive recurrent vision models. in *Neural Information Processing Systems (NeurIPS)* (2020).
43. Linsley, D., Kim, J., Veerabadran, V., Windolf, C. & Serre, T. Learning long-range spatial dependencies with horizontal gated recurrent units. in *Advances in Neural Information Processing Systems*. Vol. 31 (eds Bengio, S. *et al.*) (Curran Associates, Inc., 2018).
44. Zhou, Y., Liu, S. & Ma, Y. NeRD: Neural 3D Reflection Symmetry Detector. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15940–15949 (2021).
45. Zhang, H. *et al. Symmetry-Aware 6D Object Pose Estimation via Multitask Learning. Complex.* (2020). https://doi.org/10.1155/2020/8820500.
46. Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci.* **116**(43), 21854–21863 (2019).
47. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**(6), 974–983 (2019).
48. Clarke, A., Devereux, B. J., Randall, B. & Tyler, L. K. Predicting the time course of individual objects with MEG. *Cereb. Cortex.* **25**(10), 3602–3612 (2015).
49. Freiwald, W. A. & Tsao, D. Y. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science.* **330**(6005), 845–851 (2010).
50. Brincat, S. L. & Connor, C. E. Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron.* **49**(1), 17–24 (2006).
51. Spoerer, C., Kietzmann, T., Mehrer, J., Charest, I. & Kriegeskorte, N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Comput. Biol.* **16**, e1008215. https://doi.org/10.1371/journal.pcbi.1008215 (2020).
52. Huang, G., Liu, Z. & Weinberger, K. Q. Densely connected convolutional networks. *IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR).* **2017**, 2261–2269 (2017).
53. Chollet, F. Xception: Deep learning with depthwise separable convolutions. *IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR).* **2017**, 1800–1807 (2017).
54. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. *Inception-v4. Inception-ResNet and the Impact of Residual Connections on Learning* (AAAI, 2017).
55. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR).* **2016**, 2818–2826 (2016).
56. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR).* **2016**, 770–778 (2016).

57. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. *IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR).*, 248–255 (2009).
58. Cicconet, M., Birodkar, V., Lund, M., Werman, M. & Geiger, D. A convolutional approach to reflection symmetry (2016). http://arxiv.org/abs/1609.05257.
59. Liu, J. *et al.* Symmetry detection from realworld images competition 2013: Summary and results. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 200–205 (2013).

## Acknowledgements
The authors wish to acknowledge and thank Prof. Pawan Sinha at the Department of Brain and Cognitive Sciences at MIT for his support and guidance. This work has been supported by Fujitsu Limited (Contract No. 40009105) and the R01EY020517 Grant from the National Eye Institute (NIH).

## Author contributions
All authors designed research; S.S., D.S., M.G. performed experiments with contributions of X.B.; S.S., D.S. and X.B. wrote the paper with contributions of M.G. and T.S.; T.S. and X.B. supervised the research.

## Competing interests
This study received funding from Fujitsu Limited. The funder through TS had the following involvement with the study: conception of the experiment, writing of this article and supervision of the study. All authors declare no other competing interests.

## Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-25219-w.

**Correspondence** and requests for materials should be addressed to S.S. or X.B.

**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.