Trained deep neural network models of the ventral visual pathway encode numerosity with robustness to object and scene identity

Thomas Chapalain,[1]* Bertrand Thirion,[1] Evelyn Eger [2]*

[1] Inria, CEA, Université Paris-Saclay, France
[2] Cognitive Neuroimaging Unit, INSERM, CEA, CNRS, Université Paris-Saclay, NeuroSpin center, France

* Corresponding author contacts: thomas.chapalain@inria.fr, evelyn.eger@gmail.com

Short title: numerosity representation in object vision networks

Abstract

"Number sense", the ability to quickly estimate quantities of objects in a visual scene, is present in humans and many other animals, and has recently been demonstrated in biologically inspired vision models, even before training. However, real-world number perception requires abstraction from the properties of individual objects and their contexts, in contrast to the simplified dot patterns used in previous studies. Using novel, synthetically generated photorealistic stimuli, we discovered that deep convolutional neural networks optimized for object recognition can encode numerical information across varying object and scene identities in their distributed activity patterns. In contrast, untrained networks failed to discriminate numbers, and appeared to encode low-level visual summary statistics of scenes rather than the number of discrete objects per se. These results caution against using untrained networks to model early numerical abilities and highlight the need to use more complex stimuli to understand the mechanisms behind the brain's visual number sense.

Introduction

Numerosity, or the number of objects in an environment, is an important parameter of variation in visual scenes that we can quickly apprehend in an approximate, ratio-dependent manner. In contrast to precise counting and the use of symbols based on language, this visual perceptual "number sense" is widespread across the animal kingdom (1, 2) and is present, at least in crude form, at very early stages of life (3, 4). Nevertheless, this basic nonverbal skill has been discussed as one possible foundational capacity enabling the development of more sophisticated numerical skills in humans (5–8), making its neural underpinnings a topic of wide interest.

Neuroscientific studies have provided evidence for a critical substrate underlying number sense predominantly in the dorsal visual pathway leading up to the intraparietal cortex in the primate brain (9, 10). FMRI activity in these regions is sensitive to numerical changes in a ratio-dependent manner (11), and direct evoked activity patterns encode information about individual non-symbolic numerical quantities (12–15) with some independence from the sets' non-numerical characteristics. Moreover, responses to specific numerosities are arranged in orderly topographic maps in the parietal cortex (16, 17) with individual recording sites exhibiting Gaussian tuning curves to preferred numbers of items, similar to those observed in single neurons of macaque monkeys and other species (18–21).

Psychophysical work increasingly points to approximate numerosity being a perceptual property that is "directly sensed" within the visual system (22, 23). However, a generally accepted computational account of how the brain extracts internal representations of number from sensory responses to visual images is still lacking. While several computational studies within the field of number cognition have been dedicated to the problem of numerosity representation, some (24–26) have not focused on explaining responses to visual images. Others have demonstrated that codes for numerical quantity can emerge from unsupervised learning in hierarchical generative networks, but remained limited to highly simplified stimuli and network architectures that bear little resemblance to the brain's visual hierarchy (e.g., 27, 28).

In visual neuroscience more generally, hierarchical convolutional neural networks (HCNNs), a class of artificial neural networks used in computer vision that was initially inspired by aspects of visual neurophysiology, have emerged over the last decade as the mainstream models of high-level visual processing in the brain (29, 30). These networks, which offer versatility of application to any type of visual image and allow for comparison of their internal representations with those found in the brain, have been particularly successful when applied to mechanisms of object recognition in the ventral visual stream (e.g., 31–35).

In this context, an important recent finding was that individual artificial neurons in higher layers of such networks, after or even before training for object recognition, exhibited preferential responses to dot numbers very similar to those observed in biological neurons (36, 37). This led to the suggestion that visual number sense could be an emergent property of suitable neural network architectures without dedicated neural mechanisms being optimized for number identification per se. That number sense could emerge from the same neural machinery as the one that supports object recognition nevertheless seems counterintuitive: The notion of "four" applies in the same way to four cats as well as to four guitars, regardless of their precise visual properties and the context of their appearance. Thus, it requires generalization across the very same factors that visual systems need to be sensitive to in order to recognize objects. Given that previous studies reporting "number neurons" in HCNNs have only tested those networks with binary dot patterns, one possibility is that such findings hold only in these narrow stimulus settings and actually reflect different aspects of the stimuli than the discrete number of items per se. For example, a larger number of dots implies a more complex stimulus (as quantified by the perimetric complexity measure) and a difference in the spatial frequency distribution (38, 39). Moreover, recent work has shown that even across changes in non-numerical attributes such as the size and spacing of items, their number in dot

patterns remains closely related to aspects of low-level contrast statistics expressed in terms of Fourier magnitude (40). This implies that apparently different responses to dot number could may very simple sensory mechanisms that would unlikely be sufficient to extract number in more general and naturalistic settings.

If artificial neural networks would fail to discriminate number when they are tested with more complex and naturalistic stimuli that dissociate the number of discrete items from such low-level image statistics, then results such as those described would not reflect a sense of visual numerosity per se but simply responses to the correlated low-level statistics. If the sensory representations of networks optimized for object recognition were unable to support the simultaneous identification of objects and numbers across changes in the objects' characteristics, this could potentially explain why visual systems require separate resources for the two types of discrimination, and hence why numerosity processing has been found to be predominantly associated with the dorsal visual pathway in the primate brain. On the other hand, the finding that these two types of information can coexist in neural networks would provide evidence for a possible, and hitherto underappreciated, contribution of the ventral visual pathway to number sense.

To explore these possibilities, here we tested whether HCNNs can encode numerosity information in original synthetic photorealistic stimuli that varied in object identity and scene context in addition to number. This type of stimulus design allowed us to perform strong and well-controlled generalization tests that would not be easily feasible with fully natural real-world scenes. It also helped us to dissociate the encoding of the discrete number of items per se from the encoding of several low-level visual summary statistics that we also quantified, which would be correlated with number in simpler stimuli or in paradigms that do not include similar generalization tests. Using this novel paradigm, we examined whether numerosity can be read out from the population activity of networks optimized for object recognition and their untrained counterparts, and the degree to which such decoding can abstract from changes in object and background scene content. We also systematically investigated the extent to which numerosity decoding performance can be explained by several low-level summary statistics of the presented images. Finally, we examined number-selective artificial neurons defined as in previous work for their encoding of numerosity in our photorealistic stimuli, and compared results obtained with our photorealistic stimuli to those obtained with more conventional binary patterns.

Our findings point to a population code for numbers of objects that supports some degree of independence from object and background changes, which is not carried by a few artificial neurons with the highest selectivity for binary dot patterns as those described in previous studies, and present only in object recognition trained (but not untrained) networks.

Results

Photorealistic stimuli decorrelate number from low-level image statistics

We generated two stimulus datasets of synthetic photorealistic images comprising multiple objects and covering two different numerosity ranges. The "subitizing range" contained stimuli representing small numerosities (1, 2, 3 or 4 objects per image) and the "estimation range" larger numerosities (6, 10, 15 or 24 objects per image). For both datasets, each stimulus type is defined by the choice of an object, a background scene and a numerosity: the chosen object is inserted into the background at randomly chosen positions and views. The choice to combine objects with unrelated background scenes to obtain rich stimulus variability is an adaptation of the strategy used with single objects in several influential studies investigating object representations in HCNNs and neural responses, e.g. (31). In addition to numerosity, we varied the size and spacing of our objects

here within four levels each (Fig. 1A). This design allowed us to test the extent to which CNN representations discriminate numerosity across changes in objects, backgrounds or both, allowing for generalization across superordinate categories and finer-grained individual variation.

We characterized each stimulus image in terms of six different low-level statistics. The simplest ones correspond to luminance-related measures (mean and standard deviation of the image pixel values), which typically show some correlation with numerosity as long as there is also an increase in the total surface area of the objects, which is on average the case in our design. We also computed the energy at high spatial frequencies which has previously been proposed as a possible proxy for numerosity in dot sets (38, 39) and could account for some psychophysical findings, as well as the aggregate Fourier magnitude as introduced by (40). Finally, we quantified two measures used in the field of natural scene representation (41) that describe the distribution of local contrast in natural images (referred to as "image complexity" and "texture similarity", see Methods section for more details). In previous work these measures correlated strongly with early visual cortex activity and we were interested in the extent to which they would provide a proxy for numerosity in photo-realistic images. Fig. S1 and S2 provide an overview of our individual stimulus types in the space of these measures. The degree of variation in each measure when object number, object identity or background identity differed is summarized in Fig. 1B. Although multiple statistics showed higher values on average for larger numerosities, changes in objects and backgrounds induced variations in the statistics that were in many cases as pronounced or even more pronounced (especially for background changes). The Pearson correlation between numerosity and each of the 6 measures across all stimuli (computed separately for the subitizing / estimation range, respectively) was -0.02 / -0.01 for luminance mean, 0.08 / 0.07 for luminance standard deviation, 0.40 / 0.52 for aggregate Fourier magnitude, 0.23 / 0.28 for HSF energy, 0.08 / 0.23 for image complexity, and -0.26 / -0.23 for texture similarity (for a more detailed overview of correlations split by either object or background, see Fig. S3). These results suggest that by using analyses that leverage the high variability due to changes in objects and backgrounds on the one hand, and by comparing numerosity predictions from CNN representations with those based on the low-level statistics on the other hand, we can reveal whether there is a representation of numerosity information in CNNs that goes beyond these basic visual summary statistics.

Numerosity information in CNNs generalizes across coarse object and scene categories

To determine the extent and nature of numerosity information in CNNs, we first investigated whether the networks' feature representations allow for numerosity decoding with generalization across the broad semantic categories of the stimuli (animals vs tools, natural vs artificial backgrounds). Therefore, the internal representations of all stimuli were extracted from three common CNN architectures for object recognition, for five different convolutional layers of approximately equal relative depth across networks, here referred to as Conv1 to Conv5 (see Table S1 for details). This first step represents the encoding of our stimulus images (Fig. 2A). We then performed a multivariate decoding analysis that splits the dataset into train/test sets separating the four superordinate categories (referred to as "coarse-grained generalization", Fig. 2B). This second step corresponds to the decoding of log-numerosity from the networks' representations (Fig. 2A). We found that for this type of generalization, only trained networks, but not their untrained counterparts, allowed for better than chance prediction of the number of objects in the image, with a clear gain in performance along the hierarchy of layers, from Conv1 to Conv5 (Fig. 2C). These results point to the importance of network training for object recognition, as well as the hierarchical level of the layer representations, as two key factors facilitating the encoding of numerosity with some independence from high-level visual scene properties.

However, it is possible that the decoding of the number of objects in an image was driven by non-numerical quantities such as item or total surface area, sparsity or field area, which can never be decorrelated from numerosity all at the same time. In the way our stimulus space is designed, all four of these quantities had an equal degree of correlation with numerosity (Pearson correlation r = 0.35, see Fig. S4). This suggests that the decoding analysis may have achieved better-than-chance predictive performance by using these quantities rather than relying on a representation of the items number per se. To address this issue in a revised generalization scheme, in addition to the "coarse-grained generalization" split, we constrained the train and test stimuli so that their respective non-numerical features were made congruent (C) or incongruent (I) with numerosity (Fig. 2D). This meant that during decoder training, the item area (IA) and sparsity (Spar) of the stimuli remained constant, while their total area (TA) and field area (FA) increased with numerosity. In contrast, when the decoder was tested, the TA and FA of the stimuli were constant, while their IA and Spar decreased with increasing numerosity. With this additional constraint, networks trained on object recognition still allowed for decoding of numerosity across changes in high-level visual properties (Fig. 2E), and as in the previous analysis, the predictive power for numerosity increases along the hierarchy of layers. Compared to the simple "coarse-grained generalization", there was only a small decrease in decoding performance, with prediction errors increasing by on average 4% of the chance-level. These results show that the ability to read out numerosity from the internal activation patterns of CNNs cannot be mediated by the aforementioned non-numerical quantities. Taken together, these results suggest that when tested with our photorealistic stimuli, the number of objects is inherently encoded in object recognition trained networks, but not in their untrained counterparts, in a manner that is robust to coarse object and scene categories and classical non-numerical quantitative parameters.

Generalization across exemplars and influence of low-level image statistics

To better understand which image changes trained and untrained network representations can or cannot generalize across, we investigated what was the degree of generalization of numerosity discrimination across individual objects and backgrounds, respectively. For each object (or background), we trained a decoder to estimate the number of items across all stimuli containing that specific object (or background). We then used the trained decoder to predict numerosity for each left-out object (or background), resulting in a pattern of pairwise generalization scores (Fig. 3A). When we applied this "fine-grained generalization" decoding analysis to the CNNs' representations of the stimulus images, we observed a decrease in decoding performance when predicting numerosity across changes in objects and backgrounds compared to testing on the same condition (Fig. 3B). The cost in predictive performance with generalization was particularly pronounced for the untrained networks and in the case of background changes, which, according to our previous quantification of low-level image statistics, was the type of change that induced the most pronounced differences in these measures. The pattern of performance obtained for the untrained networks may therefore suggest that this decoding was driven by some combination of low-level image statistics rather than number per se.

To further explore this idea of reliance on low-level image statistics, we performed two additional analyses. First, we tested what level of average performance could be achieved when trying to predict numerosity from the low-level statistics themselves. Using the six measures (luminance mean: L_m, and standard deviation: L_{sd}, high spatial frequency energy: E_{high}, aggregate Fourier magnitude: M_{agg}, image complexity: S_C and texture similarity: S_T) as input features for multivariate numerosity decoding, we found that they allowed for only marginally better than chance performance in a few cases where no generalization across objects and backgrounds was involved (Fig. 3C). The numerosity predictions from the trained networks' representations clearly

and consistently outperformed the numerosity predictions from these low-level image statistics (both when each statistic was considered in isolation and when they were all used together), but the prediction performance from the untrained networks did not exceed that of the low-level statistics by much. These results show demonstrate that the generalization performance of trained networks cannot be explained by low-level image statistics alone, at least within the measures examined here. Conversely, the representations of untrained networks appear to function in a manner more similar to these simple low-level models.

Next, we reasoned that if the ability to decode numerosity from a network relied on some combination of the low-level image statistics as a shortcut, we would expect to see some relationship between the six image statistics and the numerosity predicted from the networks' representations on a trial-by-trial basis. Therefore, we tested for an explicit linear mapping between low-level statistics and numerosity predictions on left out stimuli using multiple linear regression. This regression included the six image statistics together as predictors, and estimated their effect on the predictions obtained when generalizing numerosity decoding across individual objects or backgrounds. The explained variance computed from this analysis served as a metric to assess the extent to which the selected statistics could account for numerosity predictions across changes in stimuli (Fig. 3D). We found that for both types of generalization, as well as for both numerosity ranges, the variance explained by low-level statistics differed significantly between trained and untrained networks, being higher for untrained networks (objects/subitizing: $p = 1.16 \times 10^{-3}$, objects/estimation: $p = 6.10 \times 10^{-5}$, backgrounds/subitizing: $p = 3.05 \times 10^{-4}$, backgrounds/estimation: $p = 6.10 \times 10^{-4}$, Wilcoxon signed-rank test). In particular, for generalization across backgrounds, the explained variance was nearly negligible for the trained networks, with the exception of Conv1 (Fig. 3E). This layer has been documented in the previous literature to perform basic filtering with weights similar to Gabor filters (42), suggesting some degree of functional similarity to the hierarchical random filter banks that make up the untrained networks. These results further support the hypothesis that the number of discrete objects is encoded only in trained networks, whose numerosity prediction performance not only clearly outperforms the one of the low-level models, but also is significantly less modulated by low-level statistics on a trial-by-trial basis. Numerosity decoding from the untrained networks instead substantially reflects basic visual summary statistics, which may account for the poor performance when generalizing across objects and backgrounds.

Object number information is not carried by dot pattern tuned neurons

Given that previous studies (36, 37) reported artificial neurons discriminative of numerosity in AlexNet's fifth convolutional layer (Conv5), whose tuning characteristics showed similarities to real neurons previously recorded in macaque monkeys, we wondered to what extent this type of neuron might underlie the encoding of numerosity in CNN representations when using our photorealistic stimuli. Therefore, we reproduced the analysis pipeline to identify these units. We generated the same types of dot arrays as used in the previous work, containing numerosities ranging from 1 to 30, in three stimulus sets, using different controls for non-numerical quantities such as dot size, area etc. We then tested the response of AlexNet/Conv5 artificial neurons to these stimuli to find the set of artificial neurons that showed significant selectivity for dot numbers (see Methods for more details on the definition of these units). We obtained N_s=735 number-selective neurons and N_{ns}=997 non-selective neurons for AlexNet (1.70% and 2.3% of all units, respectively) and N_s=233 and N_{ns}=155 for random AlexNet (0.54% and 0.27% of all units, respectively) in the fifth convolutional layer (Fig. 4A).

We then compared the ability of these number-selective neurons to decode numerosity with that of non-selective neurons, for both AlexNet and its untrained counterpart, when using our

previously described coarse-grained generalization approach on our photorealistic stimuli, this time on the union of the subitizing and the estimation range stimuli (Fig. 4B). Using either the set of number selective neurons or the set of non-selective neurons, for both trained and untrained AlexNet, we observed a failure to generalize: decoding performance was either at chance or worse than chance, approaching the poor performance obtained when using image pixel values which served as a baseline (Fig 4C-left). However, we found that the larger population of artificial neurons of the trained AlexNet, excluding the set of number neurons, showed successful numerosity decoding while the untrained network did not, replicating the results of the previous coarse-grained generalization analysis. This finding suggests that artificial neurons tuned to dot patterns are specific to such simplified stimuli and fail to consistently represent numerosity in the face of significant changes in high-level stimulus properties, such as changes in object and scene identity.

Finally, to further assess the influence that the number of artificial neurons used as input for decoding had on the ability to read out numerosity, we repeated the previous type of decoding analysis while randomly sampling decreasing fractions of the total population of artificial neurons. We distinguished two different regimes. First, when the fraction of units used in the analysis was reduced from the whole population to its tenth, the predictive performance remained constant, suggesting a high redundancy in the shared information across units. Then, when we further reduced the total number of units to 108 (0.25% of the full population size), we observed a strong drop in decoding performance to chance-level, indicating that very few randomly selected units of the trained AlexNet could not encode numerosity in a way that generalized across object and background category changes (Fig. 4C-right). Taken together, these results suggest that for the networks studied here, the representation of object numerosity with some tolerance to other high-level image changes appear to be better described by a robust code distributed across the population of artificial units rather than by a code carried by few number-selective units.

Simplified stimuli reduce differences between trained and untrained networks

Overall, in our investigation of the encoding of object number information in CNNs using synthetic photorealistic stimuli, we have observed important differences between object recognition trained and untrained networks that were not found in the previous literature. To confirm whether these differences were related to the use of more complex photorealistic stimuli, we next simplified our stimuli to determine the extent to which differences between the results of trained and untrained networks would vanish. To do this, we binarized each photorealistic stimulus image, setting the pixels enclosed by the objects to white, and background pixels to black, resulting in a simplified stimulus set that more closely resembled the binary dot arrays used in previous studies. When the feature representations of the CNNs for these simplified stimuli were used as input for decoding in the fine-grained generalization scheme across objects (Fig. 5A), both trained and untrained networks showed better than chance decoding performance across all layers (Fig. 5B). Testing for effects of model training (trained vs. untrained) and stimulus type (photorealistic vs. segmentation mask) in a 2-way repeated measures ANOVA revealed a significant interaction between training and stimulus type: ($F(1, 29) = 368$, $p = 5.06 \times 10^{-18}$), corresponding to a smaller difference between trained and untrained models in the case of the simplified binary images compared to the photorealistic ones. While a discernible difference in performance between the trained networks and their untrained counterparts persisted in the case of the simplified stimuli ($p = 0.016$, Wilcoxon signed-rank test), the effect size was small: $\Delta MAE = 0.02$, representing a relative difference of 9% from the average MAE performance across trained and untrained networks compared to $\Delta MAE = 23\%$ for the complex photorealistic stimuli ($p = 1.73 \times 10^{-6}$, Wilcoxon signed-rank test). These results show that the ability to extract numerosity from simplified stimuli such as binary sets of items, even with generalization across objects with pronounced shape changes, is not substantially

different between untrained and object-recognition-trained networks, similar to the pattern of findings in the previous literature investigating dot-pattern tuned units. This further confirms that the crucial factor leading to differences between the results of object recognition trained and untrained networks here is the use of the photorealistic versions of the stimuli, rather than other ways in which our stimulus design or analysis differs from previous studies.

Discussion

Using original photorealistic stimuli, in this study we investigated the ability of deep convolutional neural networks, the current dominant models of high-level visual processing in the primate brain, to represent numerosity. Specifically, we tested the extent to which such networks can encode information about the approximate number of objects with invariance to changes in object and scene identity, for both small (1 to 4) and large (6 to 24) sets of objects. We found a dissociation between the networks as a function of training: networks optimized for object recognition allowed for the number of objects in a visual scene to be read out across the aforementioned changes in high-level visual properties from distributed patterns of activity in their higher layers. Randomly initialized untrained networks, however, showed limited ability to represent numerosity information that invariance to other visual properties, especially when tested across changes in background which induced the most pronounced changes in low-level attributes in our stimuli. The failure of the untrained networks' representations to support this type of discrimination appeared to be related to the fact that their encoded information was predicted to a greater extent by a combination of low-level visual summary statistics of the images.

Our finding of a clear difference between the ability of object recognition trained and untrained networks to represent numerosity contrasts with the findings of previous studies investigating responses to numerosity in simple visual stimuli such as dot patterns, which suggested that networks such as those used here contain artificial neurons that respond preferentially to different numbers of items, both before and after object recognition training (36, 37). As noted above and in the previous literature (40), the use of dot pattern stimuli cannot distinguish between effects related to the discrete number of items per se and certain associated low-level image statistics. Our results suggest that, at least for the untrained networks, the previous findings of number-selective units, whose activity could be read out to perform numerosity discrimination, did not imply a true abstract representation of discrete numerosity, but were instead driven by low-level statistics. This also implies that, contrary to previous suggestions (37), the mere projection through a hierarchical convolutional architecture, without optimized feature weights, is not sufficient to generate abstract numerosity information. Our results also differ from previous studies in the type of the code underlying numerosity discrimination in networks trained on object recognition. We find that the information used to distinguish object number in photorealistic stimuli is not carried by few artificial neurons that are most discriminative of dot numerosity, but rather appears to be widely distributed throughout the population of artificial neurons and is coded somewhat redundantly (requiring about ten percent of randomly selected units for maximum readout performance).

In human perception, enumeration of small sets of items is typically fact and error-free, a phenomenon referred to as subitizing, but enumeration of larger numbers under conditions when counting is prevented is approximate and follows a Weber's law signature (43, 22) as long as the items are not too cluttered. We found no equivalent of the distinction between these two regimes in the networks studied here, where prediction of (the logarithm of) numerosity followed a similar approximate signature, and if anything, appeared slightly less accurate in the subitizing range. Differences in human performance in the subitizing and estimation range have often been explained by the operation of two types of cognitive systems, one dedicated to object individuation/tracking

and one dedicated to the representation of large approximate magnitudes (44, 45). Interestingly, studies using dual tasks that occupy subjects' attention have found that without the availability of spatial attentional resources, numerosity discrimination performance does indeed become approximate over the entire range (46, 47). Given that spatial attention involves recurrent/feedback processing that is not present in the purely feed-forward network architectures studied here, the present finding of approximate performance across the whole numerosity range is not surprising and is indeed consistent with the human literature.

The artificial vision community has long recognized that the representations learned by CNNs trained on object recognition can contain some information useful for other visual tasks as well (48, 49). Still, it may be somewhat surprising that object number is one of them, given that identifying objects as opposed to numbers across changes in objects or other high-level visual properties, intuitively seems to present quite orthogonal computational challenges. However, the results obtained here for number nicely complement some findings in the previous literature on the representation of objects and other types of category-orthogonal visual information: For the encoding of the size and location of individual objects, it has been observed that, although these properties are orthogonal to the task objective of an object recognition CNN, information about them is not lost but is actually enhanced or encoded more explicitly along the hierarchy of network layers, as it is also along the visual hierarchy of the brain (50). In addition, information about object pose and background appears to be encoded in distributed population activity in a way that is "factorized" and thus separated from object identity (51). In these studies, distributed population codes achieved the greatest separation between the different types of information at later visual stages. This parallels our finding for number, where the ability to read out this property across changes in object and scene category was also highest in later convolutional layers. Thus, our results provide further evidence along the same lines, showing that the representational capabilities of object recognition CNNs can be far richer than the types of discriminations on which they are explicitly trained, even when it comes to properties of sets of multiple objects.

What kind of representational primitives could underlie the ability of the networks (or visual systems in general) to encode information about both what an object is and about how many objects there are independently of their detailed properties, without important trade-offs or even incompatibilities? While our own results cannot provide a definitive answer to this question, recent studies suggested that feature selectivity not only within object recognition CNNs, but also in the brain's ventral visual pathway, is not specific to whole objects (52, 53). More specifically, high-level visual representations at these stages have been found to be "texture-like" in that they can actually be "blind" to certain image changes that alter the global object shape (53), contrary to what is typically assumed based on our perceptual experience of objects as whole entities. Rather than being a limitation, the authors suggested that these findings reflect the use of a general, flexible basis set of visual features that can provide versatility for many different types of visual discrimination.

Our results have important implications for the study of visual number sense in the human brain and behavior. First, they suggest that cognitive neuroscience research should not overlook a possible contribution of the ventral "vision for recognition" pathway to this ability. Most research in the domain of number cognition has focused on dorsal (fronto-parietal) areas, although some early findings in non-human primates suggested that a smaller percentage of number-selective neurons also exist in the inferior temporal cortex (18). The responses of these neurons, however, appeared to be less independent of non-numerical properties than those in fronto-parietal regions. More recently, similar neurons have also been reported in humans in the parahippocampal and entorhinal cortices (54), which can be considered as the highest levels of the ventral visual pathway. Moreover, recent evidence suggests that distributed fMRI activity patterns of the ventral pathway can encode information about the numerosity of dot sets independently of non-numerical (size and spacing related) parameters (55), as in previous reports for the dorsal pathway (15). While the

numerical cognition field has made extensive use of dot sets as stimuli, few recent studies have used more naturalistic stimuli (56, 57). Future studies should examine population activity in both ventral and dorsal pathways with stimuli and generalization tests similar to those used here, to understand the extent to which representations in both streams are equivalent or emphasizing complementary types of information, and which stages are most important in determining behavioral numerical discrimination performance in humans.

Individuals with pronounced difficulties in acquiring mathematical skills (in the most extreme cases referred to as developmental dyscalculia) (58) beyond symbolic numerical processing often also show impairments in basic non-verbal number sense (6, 7). Interestingly, these impairments have in some cases been found to vary with the degree to which the stimulus conditions used require the separation of discrete number from other quantities (59, 60). Dyscalculics may also exhibit more general visual impairments related to the segregation and detection of visual objects, manifested by increased visual crowding, and reduced 2D shape and 3D structure from motion sensitivity (61, 62). When investigating visual number sense using dot sets, as is commonly done, a number of strategies based on low-level summary statistics rather than representations of segregated objects would be usable to achieve some success on the task. Therefore, re-examining non-symbolic number sense in dyscalculia with more complex and variable stimuli similar to those used here may make potential difficulties more apparent if they are due to impaired visual mechanisms for extracting multiple objects.

The presence of information on dot pattern numerosity in untrained neural networks (28, 37) has previously been proposed to explain how humans or other animals can already have some sensitivity to numerosity information very early after birth, which is subsequently refined during development (6). It is commonly assumed that infants already possess relatively abstract numerical abilities, based on the fact that their numerical discrimination generalizes across sensory modalities (4, 63). However, given that some low-level statistics closely follow numerosity in spatial dot sets (40), a similar relationship may exist between the number of flashes and equivalent measures in the temporal domain, and the question is to what extent generalization might occur at this level. At least some existing evidence suggests that the processing of spatial and temporal frequencies shows some correspondence between the visual and auditory modalities, respectively (64). Therefore, cross-modal sensitivity to numerosity in previous studies need not automatically imply that young infants' visual systems would support the same type of invariant representation within the visual modality as we have tested in this study. Although some infant studies have used numerosity stimuli consisting of different objects (65–68), to our knowledge none have tested for generalization across both object and scene identities with important differences in low-level summary statistics, as we did here. If untrained networks are a sufficient model of newborns' visual abilities, then we would expect that the ability to generalize numerical discrimination across similar types of changes would not be present very early in life. It is known that although basic visual function including some degree of shape discrimination is present soon after birth, the ability to discriminate object categories continues to mature up into adolescence, see (69) for a review. The possibility that the representation of numerosity with abstractness to other high-level visual changes would also show a somewhat protracted maturation related to other high-level visual abilities still remains open.

It is of course possible that the untrained neural network with random weights is an oversimplified and incomplete model of the newborn visual system in several respects. Mammalian visual systems do already show preliminary tuning to at least some visual features at the onset of visual experience, which appears to develop under the influence of spontaneous organized waves of activity that occur in the visual pathways before birth or eye opening (70). In addition, the brain's subcortical visual pathways, which are thought to be precociously functional (71, 72), have no equivalent in the current neural network models and may be responsible for visual abilities going beyond those of the untrained networks studied here. More generally, it seems worth considering that visual number sense need not necessarily rely on a single mechanism, but that different

computations and brain substrates could contribute in different ways across development. The fact that the relative salience of changes in number, as opposed to object identity, has been found to change during early years of life (68) could be consistent with the involvement of multiple visual mechanisms that are active at different timepoints.

The numerical information observed here in artificial neural networks appeared to be widely distributed across the population of artificial neurons, but could be read out by a linear decoder trained to discriminate numerosity in a supervised fashion. This raises the question of how such information could be exploited in biological visual systems without the availability of explicit numerical labels, i.e., before learning to count. The fact that we have used a supervised decoder here does not necessarily mean that training with explicit numerical labels is strictly necessary for higher areas to be able to extract this information. Indeed, some recent modeling work interestingly suggests that a "number-line-like" representation of numerical magnitude can emerge at higher level of a neural network after observing the changes in visual images when items are added, removed or just changed in position (73), at least for simplified dot sets. This suggests that feedback from action systems during spontaneous manipulation of objects and observation of the resulting visual changes could play an important role in the development or refinement of number sense.

The networks in which we observed numerosity information across changes in object and scene identity here were optimized for object recognition, which is the task that has made such networks widely popular in visual neuroscience research, and also the type of CNN that has previously been investigated in the context of number sense research. This does not mean that the findings obtained are necessarily specific to object recognition training per se, but it may be that numerical information in deep convolutional architectures arises in the context of experience with natural images in general, and is only modulated in degree by the demands of the trained task. An important goal for the future will be to investigate how the representations in object recognition networks differ from those resulting from optimization for other visual task goals, some of which might be hypothesized to be more relevant to the dorsal visual pathway's general task goal of object-directed action. Comparing the representations of networks that differ in task goal and architecture in terms of their ability to account for brain representations in the dorsal and ventral visual pathways could provide important insights into the constraints that shape the emergence of number sense in the brain.

Materials and Methods

Experimental Design

Stimulus generation for this study employed a set of twenty 3D object models (half animals, half tools) obtained from https://www.turbosquid.com/. Using Blender, a free and open-source 3D computer graphics software (https://www.blender.org/) and its Cycles rendering engine, photorealistic renderings of each 3D object model were generated from 48 viewpoints, The viewpoints are described via the standard unitary spherical coordinates (r, θ, ϕ) [deg], where the 3D model corresponds to the center of the sphere, $\theta \in \{-45, -15, 15, 45\}$, and $\phi \in \{+/-15, +/-45, +/-75, +/-105, +/-135, +/-165\}$. The reference angle (θ, ϕ) = (0, 0) was determined individually for each object. Specifically, these reference angles were established in an as far as possible semantically meaningful way, to ensure a reasonable degree of consistency across all objects. For instance, animals were consistently oriented with their heads upright and facing forward, while tools were positioned in alignment with how they are typically encountered in everyday life, with their primary rotational symmetry oriented along the vertical axis, if they possessed one. High Dynamic Range Images (HDRI) obtained from www.polyhaven.com were used to create the environmental

background scene, which was also used to create naturalistic lighting for the object. Each model was rendered with the lighting corresponding to 20 different scenes (half natural, half artificial), in addition to rendering a view of each background scene separately. Randomly drawn views of each given object were then inserted into each one of the background scene renderings (see Fig. 1), resulting in a total of 400 object-background combinations. The stimuli contained 1,2,3 or 4 objects for the small (subitizing) numerical range and 6,10,15 or 24 objects for the large (estimation) numerical range. The stimuli also varied in non-numerical quantities (4 levels of size and spacing, see Fig 1 and Fig. S4 for additional details) similar to the approach used in previous work (74). By creating two instances (with randomly chosen stimulus positions and views) for each set of parameters (numbers, sizes, spacings, and object-background combinations), our semi-realistic stimulus dataset comprised a grand total of 51 200 stimuli for each of the two numerosity ranges (or 12 800 per number condition).

Characterization of Low-Level Image Statistics

For all stimuli we quantified several low-level image statistics related to luminance, spatial frequency or local contrast content:

Luminance represents the grayscale value, with L = 0 corresponding to black and L = 255 to a white pixel. We computed the mean of the luminance L_m and its standard deviation L_{sd} which represents the RMS contrast of the stimulus. By denoting L_{ij} the luminance of the (i, j) th pixel of a stimulus of size H × W, L_m and L_{sd} correspond to:

$$L_m = \frac{1}{HW}\sum_i\sum_j L_{ij} \qquad \text{and} \qquad L_{sd} = \sqrt{\frac{1}{HW}\sum_i\sum_j \left(L_{ij} - L_m\right)^2} \qquad (1)$$

For the original 900-pixel images, the frequency range spans from ν_{min} = 1 cycle/image to ν_{max} = 450 cycles/image. We computed E_{high}, the total rotational average of the energy in the high spatial frequency range, by using a cut-off frequency (ν_{cut}) between the low and high spatial frequencies. This cut-off frequency corresponded to a half-cycle matching the average object size: 184 pixels for the subitizing range and 75 pixels for the estimation range. By denoting $F(\nu_x, \nu_y)$ the Fast Fourier Transform (FFT) of the stimulus at the spatial frequency (ν_x, ν_y) and $E_r(\nu)$ its rotational average energy density spectrum at frequency ν, E_{high} can be expressed as:

$$E_r(\nu) = \frac{1}{\#\{(\nu_x,\nu_y)/\nu = \sqrt{\nu_x^2 + \nu_y^2}\}} \sum_{\nu=\sqrt{\nu_x^2+\nu_y^2}} \left|F(\nu_x,\nu_y)\right|^2 \quad \text{and} \quad E_{high} = \sum_{\nu>\nu_{cut}} E_r(\nu) \qquad (2)$$

We also computed the aggregate rotational Fourier magnitude M_{agg}, generalized from its definition in (40). M_{agg} corresponds to the sum of the Fourier magnitude at all orientations and spatial frequencies up to the first harmonic frequency ν_1, which is a prominent characteristic of the FFT of a dot set. However, in the case of photo-realistic stimuli which use real views of 3D objects instead of dots, the boundaries of the harmonics do not manifest as clear throughs in the Fourier magnitude spectrum. In order to obtain a measure coming as close as possible to that used in the aforementioned previous work, we took advantage of the fact that ν_1 depends on the size of the objects, as shown in (40), and we defined an equivalent cut-off frequency $\widetilde{\nu_1}$ that would correspond to the first harmonic frequency of a circle whose area is equal to the average area of the objects inserted in the photorealistic stimulus: The FFT of a circle of radius R is given by $J_1(\nu R)/\nu$ where

$J_1(.)$ is the first order Bessel function. If we denote IA_m the average area of the objects and R_e the equivalent radius of a circle covering the same area, a surrogate for this first harmonic frequency $\widetilde{\nu_1}$ is given by:

$$\widetilde{\nu_1} = \frac{\vartheta_0}{R_e} \qquad \text{with} \qquad R_e = \sqrt{\frac{IA_m}{\pi}} \tag{3}$$

Where ϑ_0 corresponds to the first root of J_1. A valid approximation of $\vartheta_0 \approx 3.83$ can be found in numerical tables. Subsequently, the aggregate rotational Fourier magnitude M_{agg} can be derived as:

$$M_r(\nu) = \sum_{\nu=\sqrt{\nu_x^2+\nu_y^2}} \left|F(\nu_x,\nu_y)\right| \qquad \text{and} \qquad M_{agg} = \sum_{\nu<\widetilde{\nu_1}} M_r(\nu) \tag{4}$$

Measures of the "image complexity" S_C and the "texture similarity" S_T were obtained from a Weibull fit of the local contrast distribution of our stimuli S similar to (41). We first obtained the local contrast C(i, j) at pixel (i, j) of the stimulus by filtering it with a derivative of a Gaussian kernel G_σ, using a standard deviation of the Gaussian kernel $\sigma = 12$.

$$\mathrm{C(i,j)} = \sqrt{\left(S * \frac{\partial G_\sigma}{\partial \mathrm{i}}(i,j)\right)^2 + \left(S * \frac{\partial G_\sigma}{\partial \mathrm{j}}(i,j)\right)^2} \quad \text{with} \quad G_\sigma(\mathrm{i,j}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{i^2+j^2}{2\sigma^2}\right) \tag{5}$$

Then, we fitted the local contrast distribution C(c) with a Weibull distribution W of parameters (β, γ), using maximum likelihood estimation:

$$\mathcal{W}(\mathrm{c}|\beta,\gamma) = \frac{\gamma}{\beta}\left(\frac{\mathrm{c}}{\beta}\right)^{\gamma-1}\exp[-(\mathrm{c}/\beta)^\gamma] \qquad \text{and} \qquad (\hat{\beta},\hat{\gamma}) = \arg\max_{(\beta,\gamma)} \prod_c \mathcal{W}(c|\beta,\gamma) \tag{6}$$

The stimulus representation of natural images in the space of the Weibull parameters (β,γ) is characterized by two main directions of variation, see also (41). The first eigenvalue axis, corresponding to an increase in both β and γ, represents an increase in image complexity. The second eigenvalue axis shows a transition between images that have more or less homogeneous texture over the entire image. Accordingly, we obtain the "image complexity" and "texture similarity" measures of each stimulus by projecting its estimated $(\hat{\beta},\hat{\gamma})$ values onto the first and second eigenvalue axes of the Weibull parameter space of our entire stimulus set, respectively.

Artificial neural networks and Feature Extraction

Three convolutional neural network (CNN) architectures, commonly used for object classification, were examined: AlexNet, ResNet50, and VGG16, all as implemented in PyTorch 2.0. (https://pytorch.org/). We used each of these CNNs in two instances: one pre-trained on the ImageNet dataset (75) and another untrained. The weights for the untrained models were randomly initialized using the initialization method described in (76). For our analysis, we focused on the post-ReLU activations of five convolutional layers located at comparable depth relative to the network's size. A detailed overview of these layers is provided in Table S1. Stimulus images were

down-sampled to the input resolution of the networks (224 x 224 pixels) and then fed to each of the selected networks to obtain their successive feature representations along the hierarchy of convolutional layers.

Numerosity Decoding Analysis

We used linear Ridge regression as implemented in scikit-learn (https://scikit-learn.org/stable/) to predict the logarithm of numerosity from the CNN feature representations of the stimuli and evaluated the decoding performance of each network on left-out stimuli using the mean absolute error (MAE) metric. Perfect decoding performance would result in a MAE score of 0. The optimal ridge penalization hyperparameter λ was selected by a nested leave-one-sample-out cross-validation loop during the training phase. To determine the chance decoding performance to be used as a baseline, the same Ridge regression was fitted on a random permutation of the feature representations of the stimuli and their associated log-numerosity targets. In addition, to mitigate potential differences in statistical power arising from unequal dimensionality of the feature representations across networks and layers, we equated the dimension of each network's feature space through univariate feature selection: one-way ANOVA feature selection for any difference between numerosities was performed during the training phase to reduce the dimension of each representation down to the smallest feature space size (i.e., 43 264 dimensions, AlexNet/Conv5) and subsequently applied during the testing phase.

In the coarse-grained generalization scheme, the stimulus dataset was split into train/test sets separating the four superordinate categories resulting in four subspaces, each consisting of 10 objects paired with 10 backgrounds (e.g., animals on natural backgrounds as training set and tools on artificial backgrounds as test set). For each of the 4 possible train-test set combinations, we performed a 6-fold cross-validation where in each cycle, we randomly selected 20 object/background conditions from the training set, resulting in a total of 1280 samples to learn the Ridge regression parameters. Within each fold, the MAE scores for this trained Ridge regression were then calculated using the entire test set. Finally, all the MAE scores were averaged across folds and the 4 train-test set combinations to obtain the cross-validated decoding performance score for a given network layer.

In the fine-grained generalization scheme, for each single object or background, a linear Ridge regression was trained on the CNN feature representations of all stimuli containing that specific object or background, respectively. The trained decoder was then tested on either the same or each of the other left-out objects or backgrounds, resulting in generalization MAE scores between pairs of objects or backgrounds, respectively. This entire procedure was repeated twice using the two instances of each stimulus type that differed only in the views and locations where the objects were placed (see Experimental design). Across these two folds of the cross-validation, one instance of the stimulus was used for training and the second one was used for testing. Finally, we averaged the MAE scores obtained for each pair of objects or backgrounds across the two folds to derive the cross-validated pairwise pattern of generalization performance scores for the model layer. This scheme allowed us to test the predictive performance when testing on the same condition in exactly the same way as when testing on a different condition.

Analyses of Number Selective Neurons

To determine whether the results obtained for our photorealistic stimuli were based on the same type of units tuned to binary dot sets as observed in previous studies (36, 37), we first replicated their experimental approach. Using a Python translation of the code from (37), we generated a dataset of visual dot arrays, consisting of images of size 224 × 224 pixels with n = 1, 2, 4, 6, . . . 30 dots pasted at random locations, further split into three different subsets controlling for different

non-numerical geometric cues. In the first subset (standard condition), all the dots have nearly identical radii, drawn from the normal distribution N ($\mu = 7$, $\sigma = 0.7$). In the second subset (control set 1), the total area covered by the dots remains constant across numerosities (1200 pixels), and the average distance between neighboring dots is constrained within a narrow range (90 to 100 pixels). In the third subset (control set 2), the convex hull formed by the dots is fixed (a pentagon with a circumference of 647 pixels) for numerosities greater than 4, and the individual dots vary randomly in shape between circle, rectangle, ellipse, and triangle.

The preferred numerosity (PN) of an artificial numerosity-selective neuron was defined as the numerosity that elicited the largest response on average among the responses for all dot array stimuli. To determine the average tuning curves of all number neurons, the tuning curves of all units were normalized between 0 and 1 and then averaged across neurons with the same PN value.

We identified units as number selective or non-selective using two-way ANOVA procedures and thresholds as those used in the previous studies (also see statistical analysis section below). Because it has been suggested that the detection of number-selective neurons may be unreliable with small ANOVA sample sizes (77), we used 900 images for each numerosity and stimulus set, resulting in a total of $900 \times 16 \times 3 = 43\,200$ images, comparable to the number of units in the network layer analyzed (AlexNet conv5).

Finally, we evaluated the decoding performance based on the numerosity-selective neurons compared to the non-selective neurons in the coarse-grained generalization analysis on our dataset of photorealistic stimuli. For comparison, and to assess whether the numerosity information in the CNNs comes solely from the number neurons, we also performed this analysis with the entire population of artificial units, but without the number-selective neurons. Unlike our previous analyses, here, the stimuli from both subitizing and estimation ranges were combined into a unified training and testing set to decode numerosity $N \in \{1, 2, 4, 6, 10, 24\}$. Finally, to further evaluate whether the number of units used for the decoding affected numerosity prediction performance, we performed the same decoding analysis by randomly sampling different fractions of the total artificial unit population. For each percentage p of the population retained for the decoding, we bootstrapped the procedure $\lfloor 100/p \rfloor + 1$ times, allowing for statistical sampling of the entire population of units across the iterations.

Mapping low-level statistics to the fine-grained numerosity predictions

To assess the extent to which the selected low-level statistics could account for the networks' numerosity predictions across stimuli, we used multiple linear regression with the six image statistics together as predictors and the predicted log-numerosity from the fine-grained generalization across objects or backgrounds, respectively, as the target. We performed this analysis for all the different training conditions of the fine-grained generalization scheme (20 objects & 20 backgrounds). This amounts to fitting a specific multiple linear regression on the centered 20 x 2 x 1280 = 51 200 log-numerosities predicted for the test images left out during training (19 objects/backgrounds, and the version of the same object/background not seen during training) and the standardized low-level statistics values of these images concatenated together. For each trained object or background, the explained variance (R^2) of the regression model was computed, indicating how well the image statistics accounted for the variability in the predicted log-numerosity. We then averaged the explained variance across all the 20 fits, yielding the final average explained variance score for the generalization across objects and backgrounds, respectively.

Statistical analysis

All statistical variables, including the sample sizes, exact p-values, and statistical methods, are reported in the corresponding texts or figure legends. Repeated-measures ANOVA and two-tailed

Wilcoxon signed-rank tests were used for analyses comparing between predictive performance of different models. The Wilcoxon signed-rank tests correspond to the paired version of this test, by first computing the difference between trained and untrained predictive MAE as input (corresponding to a 15-sample test for Fig. 3 and 30 samples for Fig. 5). Pearson's r was used for all reported correlations.

Analysis of number-selective units used two-way ANOVA with numerosity and stimulus sets as factors. Network units that showed a significant main effect of numerosity ($p < 0.01$), but no significant effect of stimulus set or interaction between factors ($p > 0.01$) were labeled as number neurons. Conversely, units that showed a significant main effect of stimulus set ($p < 0.01$) but no significant effect of numerosity or the interaction between factors ($p > 0.01$) were labeled as non-selective neurons.

References

1. A. Nieder, Neuroethology of number sense across the animal kingdom. *Journal of Experimental Biology* **224**, jeb218289 (2021).

2. E. Lorenzi, M. Perrino, G. Vallortigara, Numerosities and Other Magnitudes in the Brains: A Comparative View. *Front Psychol* **12**, 641994 (2021).

3. F. Xu, E. S. Spelke, Large number discrimination in 6-month-old infants. *Cognition* **74**, B1–B11 (2000).

4. V. Izard, C. Sann, E. S. Spelke, A. Streri, Newborn infants perceive abstract numbers. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10382–10385 (2009).

5. J. Halberda, M. M. M. Mazzocco, L. Feigenson, Individual differences in non-verbal number acuity correlate with maths achievement. *Nature* **455**, 665–668 (2008).

6. M. Piazza, *et al.*, Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition* **116**, 33–41 (2010).

7. M. Schneider, *et al.*, Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: a meta-analysis. *Dev Sci* **20**, e12372 (2017).

8. G. Decarli, *et al.*, Severe Developmental Dyscalculia Is Characterized by Core Deficits in Both Symbolic and Nonsymbolic Number Sense. *Psychol Sci* **34**, 8–21 (2022).

9. E. Eger, "Neuronal foundations of human numerical representations" in *Progress in Brain Research*, The Mathematical Brain Across the Lifespan., M. Cappelletti, W. Fias, Eds. (Elsevier, 2016), pp. 1–27.

10. A. Nieder, S. Dehaene, Representation of number in the brain. *Annu. Rev. Neurosci.* **32**, 185–208 (2009).

11. M. Piazza, V. Izard, P. Pinel, D. Le Bihan, S. Dehaene, Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron* **44**, 547–555 (2004).

12. E. Eger, *et al.*, Deciphering cortical number coding from human brain activity patterns. *Curr. Biol.* **19**, 1608–1615 (2009).

13. J. Bulthé, B. De Smedt, H. P. Op de Beeck, Visual number beats abstract numerical magnitude: format-dependent representation of Arabic digits and dot patterns in human parietal cortex. *J Cogn Neurosci* **27**, 1376–1387 (2015).

14. S. Cavdaroglu, A. Knops, Evidence for a Posterior Parietal Cortex Contribution to Spatial but not Temporal Numerosity Perception. *Cereb. Cortex* (2018).

15. E. Castaldi, M. Piazza, S. Dehaene, A. Vignaud, E. Eger, Attentional amplification of neural codes for number independent of other quantities along the dorsal visual stream. *Elife* **8**, e45160 (2019).

16. B. M. Harvey, B. P. Klein, N. Petridou, S. O. Dumoulin, Topographic representation of numerosity in the human parietal cortex. *Science* **341**, 1123–1126 (2013).

17. Y. Cai, *et al.*, Topographic numerosity maps cover subitizing and estimation ranges. *Nat Commun* **12**, 3374 (2021).

18. A. Nieder, E. K. Miller, A parieto-frontal network for visual numerical information in the monkey. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7457–7462 (2004).

19. P. Viswanathan, A. Nieder, Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices. *Proc Natl Acad Sci U S A* **110**, 11187–11192 (2013).

20. H. M. Ditz, A. Nieder, Neurons selective to the number of visual items in the corvid songbird endbrain. *Proc Natl Acad Sci U S A* **112**, 7827–7832 (2015).

21. D. Kobylkov, U. Mayer, M. Zanon, G. Vallortigara, Number neurons in the nidopallium of young domestic chicks. *Proc Natl Acad Sci U S A* **119**, e2201039119 (2022).

22. G. Anobile, G. M. Cicchini, D. C. Burr, Number As a Primary Perceptual Attribute: A Review. *Perception* **45**, 5–31 (2016).

23. G. M. Cicchini, G. Anobile, D. C. Burr, Spontaneous perception of numerosity in humans. *Nat Commun* **7**, 12536 (2016).

24. S. Dehaene, J. P. Changeux, Development of elementary numerical abilities: a neuronal model. *J Cogn Neurosci* **5**, 390–407 (1993).

25. T. Verguts, W. Fias, Representation of number in animals and humans: a neural model. *J Cogn Neurosci* **16**, 1493–1504 (2004).

26. T. Hannagan, A. Nieder, P. Viswanathan, S. Dehaene, A random-matrix theory of the number sense. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **373**, 20170253 (2017).

27. I. Stoianov, M. Zorzi, Emergence of a "visual number sense" in hierarchical generative models. *Nat. Neurosci.* **15**, 194–196 (2012).

28. A. Testolin, W. Y. Zou, J. L. McClelland, Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Dev Sci* **23**, e12940 (2020).

29. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

30. G. Lindsay, Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *J Cogn Neurosci* 1–15 (2020).

31. D. L. K. Yamins, *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).

32. S.-M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).

33. J. Kubilius, S. Bracci, H. P. Op de Beeck, Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Comput. Biol.* **12**, e1004896 (2016).

34. M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage* **152**, 184–194 (2017).

35. R. Rajalingham, *et al.*, Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *J. Neurosci.* **38**, 7255–7269 (2018).

36. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Sci Adv* **5**, eaav7903 (2019).

37. G. Kim, J. Jang, S. Baek, M. Song, S.-B. Paik, Visual number sense in untrained deep neural networks. *Science Advances* **7**, eabd6127 (2021).

38. S. C. Dakin, M. S. Tibber, J. A. Greenwood, F. A. A. Kingdom, M. J. Morgan, A common visual metric for approximate number and density. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19552–19557 (2011).

39. M. J. Morgan, S. Raphael, M. S. Tibber, S. C. Dakin, A texture-processing model of the "visual sense of number." *Proc. Biol. Sci.* **281**, 20141137 (2014).

40. J. M. Paul, M. van Ackooij, T. C. Ten Cate, B. M. Harvey, Numerosity tuning in human association cortices and local image contrast representations in early visual cortex. *Nat Commun* **13**, 1340 (2022).

41. H. S. Scholte, S. Ghebreab, L. Waldorp, A. W. M. Smeulders, V. A. F. Lamme, Brain responses strongly correlate with Weibull image statistics when processing natural images. *J Vis* **9**, 29.1–15 (2009).

42. M. D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, Eds. (Springer International Publishing, 2014), pp. 818–833.

43. S. K. Revkin, M. Piazza, V. Izard, L. Cohen, S. Dehaene, Does subitizing reflect numerical estimation? *Psychol Sci* **19**, 607–614 (2008).

44. L. Feigenson, S. Dehaene, E. Spelke, Core systems of number. *Trends Cogn. Sci. (Regul. Ed.)* **8**, 307–314 (2004).

45. M. Piazza, A. Fumarola, A. Chinello, D. Melcher, Subitizing reflects visuo-spatial object individuation capacity. *Cognition* **121**, 147–153 (2011).

46. P. Vetter, B. Butterworth, B. Bahrami, Modulating attentional load affects numerosity estimation: evidence against a pre-attentive subitizing mechanism. *PLoS ONE* **3**, e3269 (2008).

47. D. C. Burr, M. Turi, G. Anobile, Subitizing but not estimation of numerosity requires attentional resources. *J Vis* **10**, 20 (2010).

48. J. Donahue, *et al.*, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. [Preprint] (2013). Available at: http://arxiv.org/abs/1310.1531.

49. A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. *arXiv:1403.6382 [cs]* (2014).

50. H. Hong, D. L. K. Yamins, N. J. Majaj, J. J. DiCarlo, Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–22 (2016).

51. J. W. Lindsey, E. B. Issa, Factorized visual representations in the primate visual system and deep neural networks. *eLife* **13**, RP91685 (2024).

52. N. Baker, H. Lu, G. Erlikhman, P. J. Kellman, Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).

53. A. V. Jagadeesh, J. L. Gardner, Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences* **119**, e2115302119 (2022).

54. E. F. Kutter, J. Bostroem, C. E. Elger, F. Mormann, A. Nieder, Single Neurons in the Human Brain Encode Numbers. *Neuron* **100**, 753-761.e4 (2018).

55. A. Karami, E. Castaldi, E. Eger, M. Piazza, Neural codes for visual numerosity independent of other quantities are present both in the dorsal and in the ventral stream of the human brain. [Preprint] (2023). Available at: https://www.biorxiv.org/content/10.1101/2023.12.18.571155v1.

56. D. Odic, D. M. Oppenheimer, Visual numerosity perception shows no advantage in real-world scenes compared to artificial displays. *Cognition* **230**, 105291 (2023).

57. S. Hofstetter, S. O. Dumoulin, Assessing the ecological validity of numerosity-selective neuronal populations with real-world natural scenes. *iScience* 105267 (2022). https://doi.org/10.1016/j.isci.2022.105267.

58. B. Butterworth, S. Varma, D. Laurillard, Dyscalculia: from brain to education. *Science* **332**, 1049–1053 (2011).

59. S. Bugden, D. Ansari, Probing the nature of deficits in the “Approximate Number System” in children with persistent Developmental Dyscalculia. *Dev Sci* **19**, 817–833 (2016).

60. E. Castaldi, A. Mirassou, S. Dehaene, M. Piazza, E. Eger, Asymmetrical interference between number and item size perception provides evidence for a domain specific impairment in dyscalculia. *PLoS ONE* **13**, e0209256 (2018).

61. E. Castaldi, M. Turi, S. Gassama, M. Piazza, E. Eger, Excessive visual crowding effects in developmental dyscalculia. *J Vis* **20**, 7 (2020).

62. E. Castaldi, M. Turi, G. M. Cicchini, S. Gassama, E. Eger, Reduced 2D form coherence and 3D structure from motion sensitivity in developmental dyscalculia. *Neuropsychologia* **166**, 108140 (2022).

63. G. Gennari, S. Dehaene, C. Valera, G. Dehaene-Lambertz, Spontaneous supra-modal encoding of number in the infant brain. *Curr Biol* **33**, 1906-1915.e6 (2023).

64. E. Guzman-Martinez, L. Ortega, M. Grabowecky, J. Mossbridge, S. Suzuki, Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Curr Biol* **22**, 383–388 (2012).

65. M. S. Strauss, L. E. Curtis, Infant perception of numerosity. *Child Dev* **52**, 1146–1152 (1981).

66. P. Starkey, E. S. Spelke, R. Gelman, Numerical abstraction by human infants. *Cognition* **36**, 97–127 (1990).

67. V. Izard, G. Dehaene-Lambertz, S. Dehaene, Distinct cerebral pathways for object identity and number in human infants. *PLoS Biol.* **6**, e11 (2008).

68. C. C. Brez, J. Colombo, L. Cohen, Infants' integration of featural and numerical information. *Infant Behav Dev* **35**, 705–710 (2012).

69. V. Ayzenberg, M. Behrmann, Development of visual object recognition. *Nat Rev Psychol* **3**, 73–90 (2024).

70. A. D. Huberman, M. B. Feller, B. Chapman, Mechanisms underlying development of visual maps and receptive fields. *Annu Rev Neurosci* **31**, 479–509 (2008).

71. G. Bronson, The Postnatal Growth of Visual Capacity. *Child Development* **45**, 873–890 (1974).

72. J. Atkinson, Human visual development over the first 6 months of life. A review and a hypothesis. *Hum Neurobiol* **3**, 61–74 (1984).

73. N. Kondapaneni, P. Perona, A number sense as an emergent property of the manipulating brain. *Sci Rep* **14**, 6858 (2024).

74. N. K. DeWind, G. K. Adams, M. L. Platt, E. M. Brannon, Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition* **142**, 247–265 (2015).

75. J. Deng, *et al.*, ImageNet: A large-scale hierarchical image database in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), pp. 248–255.

76. Y. A. LeCun, L. Bottou, G. B. Orr, K.-R. Müller, "Efficient BackProp" in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, K.-R. Müller, Eds. (Springer, 2012), pp. 9–48.

77. X. Zhang, X. Wu, On Numerosity of Deep Neural Networks. *arXiv:2011.08674 [cs]* (2020).

Acknowledgments

Funding: TS was supported by a PhD stipend from Ecole Normale Supérieure.

Author contributions:

Conceptualization: EE, TS, BT
Methodology: TS, EE, BT
Investigation: TS
Visualization: TS
Supervision: EE, BT
Writing—original draft: TS, EE
Writing—review and editing: TS, EE, BT

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: The stimulus data set and analysis code will be made publicly available upon publication of the paper.

Figures and Tables

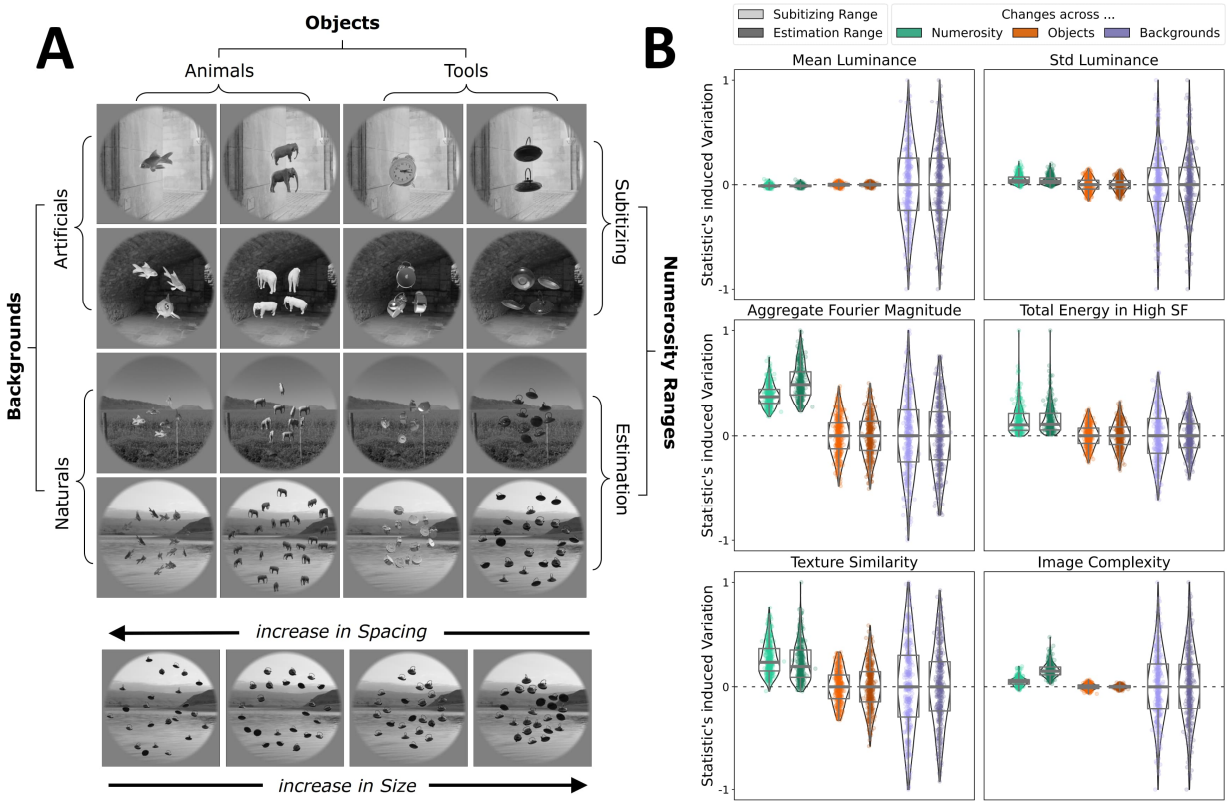

Fig. 1. Stimulus design and quantification of image statistics. (**A**) Examples of synthetic photorealistic stimuli. The images vary in terms of object (two superordinate categories, animals or tools) and background scenes (two superordinate categories, natural or artificial). They contain different numbers of objects covering either small (subitizing: 1, 2, 3, 4) or large (estimation: 6, 10, 15, 24) numerical ranges. The objects are further varied in non-numerical quantities (4 levels of size and spacing) similar to previous work (74). **(B)** Characterization of the stimulus dataset in terms of low-level summary statistics. Each panel shows how a specific low-level statistic varies when the stimulus numerosity (difference between largest and smallest), object or background (all pairwise differences) are changed. The variation is computed separately for each numerosity range (brighter or darker shade). For clarity, the plotted variation is normalized by its maximum absolute value. The whole distribution of each statistic's variations is plotted as points, with the box indicating the median, first and third quartiles.

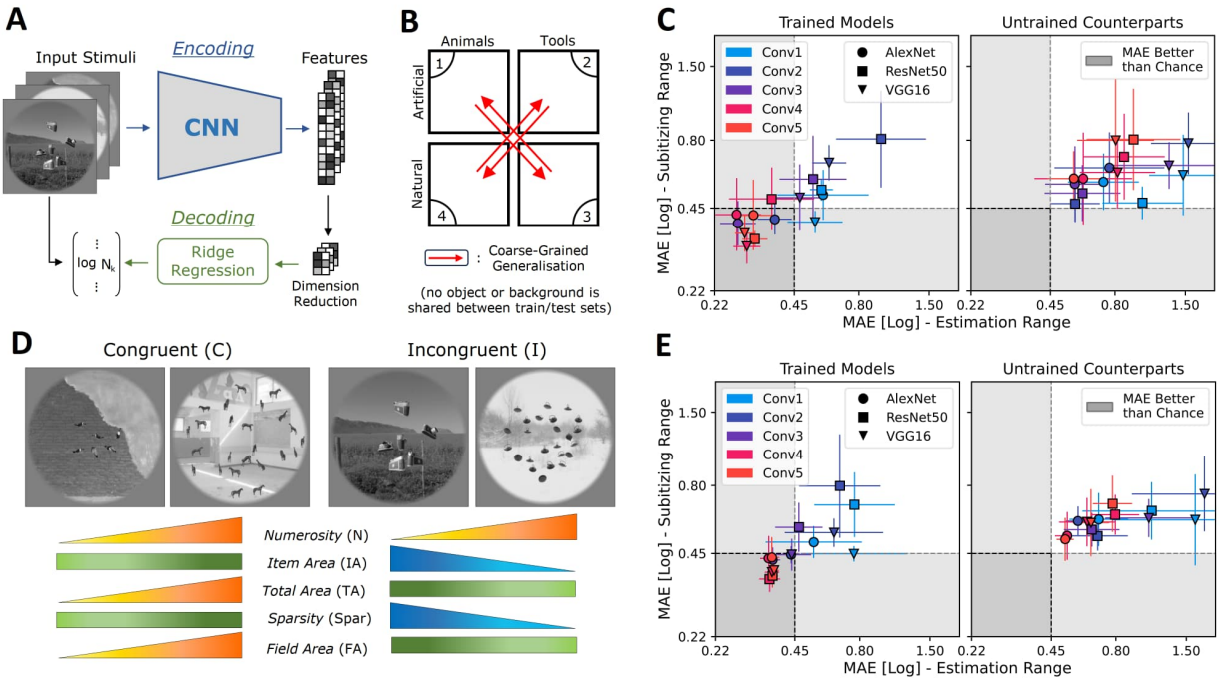

Fig. 2. Numerosity information across coarse changes in visual properties.
(A) Overview of analysis procedures. Encoding step: stimuli are fed to the CNN models. Their feature representations are extracted for several convolutional layers (Conv1 to Conv5) and undergo a feature selection. Decoding Step: Ridge regression is used to learn a linear mapping between the CNN representations of the stimuli and the logarithm of their numerosity. **(B)** Coarse-Grained generalization scheme: Decoders are trained and tested across different quadrants of the stimulus space corresponding to coarse semantic categories (red arrows). The performance is averaged for the predictive scores across all four train-test combinations. **(C)** Numerosity prediction performance (mean absolute error, MAE) in the coarse-grained generalization scheme for the different CNN architectures (markers) and layers (colors). Perfect performance is 0 and simulated chance performance is indicated by a dashed line. Error bars represent STD of the MAE over all the train-test combinations and cross-validation iterations. The subitizing range (y-axis) corresponds to 1 to 4 objects whereas the estimation range (x-axis) corresponds to 6 to 24 objects. **(D)** Stimulus subsets to control for effects of non-numerical quantities. Stimuli from each quadrant are divided into a congruent and an incongruent subset within which the non-numerical parameters (IA, TA, FA and Spar) are either increasing or decreasing with numerosity, respectively. Coarse-grained generalization is then tested between congruent and incongruent subsets. **(E)** numerosity decoding performance (MAE) for the different models on stimulus subsets from **(D)**, displayed as in **(C)**.

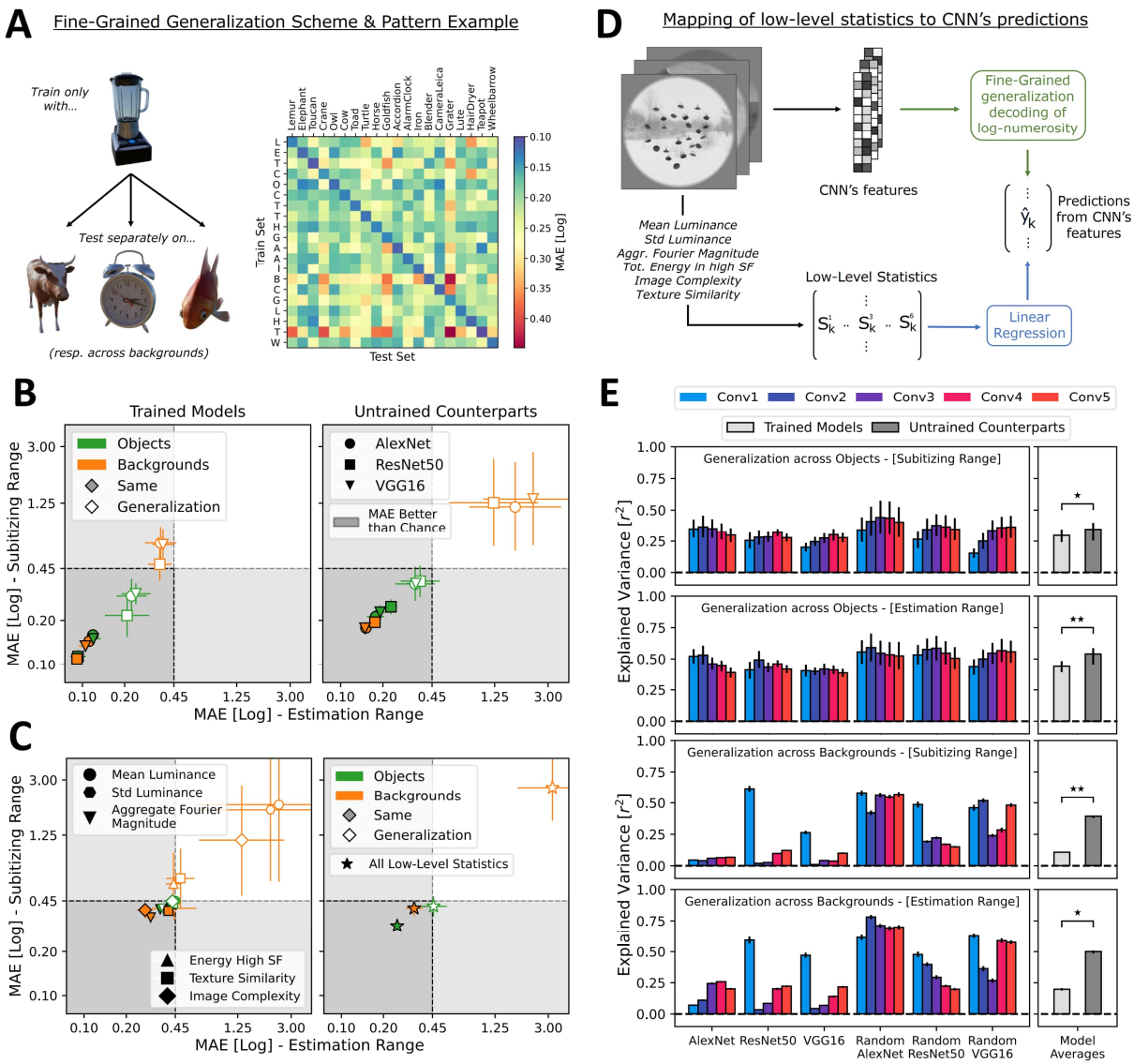

Fig. 3. Numerosity decoding across fine changes in visual properties. (A) Fine-Grained generalization scheme: For each single object (or background), Ridge regression is trained to predict the number of objects across all stimuli containing this specific object (or background). The predictive performance is computed on each left-out object (or background) using MAE, leading to a pattern of generalization scores (an example of generalization across objects is given as a 20 × 20 matrix). **(B)** Numerosity prediction performance (MAE) for the different CNN architectures (marker shapes) for the fine-grained generalization across objects (green) or backgrounds (orange). Hollow markers represent performance across pairs of different objects (or backgrounds) and filled markers represent performance for test on the same condition (means and STD across all the train-test pairs). Performance is plotted for the most predictive layer of each model (see Table S2 for the corresponding layer labels). Perfect performance is 0 and simulated chance performance is indicated by a dashed line. The subitizing range (y-axis) corresponds to 1 to 4 objects whereas the estimation range (x-axis) corresponds to 6 to 24 objects. **(C)** Average numerosity decoding performance (MAE) of the Fine-Grained generalization for several simplistic models (marker shapes) using low-level image

statistics as features for the Ridge regression. Decoding performance using all low-level statistics as a single concatenated input is shown in the right panel. **(D)** Analysis of effect of low-level statistics on predictions from CNNs: For every stimulus, six low-level statistics are computed and then mapped to numerosity predictions from the fine-grained generalization using multiple linear regression, yielding an explained variance score. **(E)** Contribution of the low-level statistics to numerosity predictions for the different CNN architectures and layers (colored bars), means and STD of the explained variance over all fine-grained generalization pairs. Explained variance averages across the trained (light gray) and untrained (dark gray) models are shown on the right. Significance levels for the difference correspond to Wilcoxon signed-rank tests: *$p < 0.005$, **$p < 0.0005$.

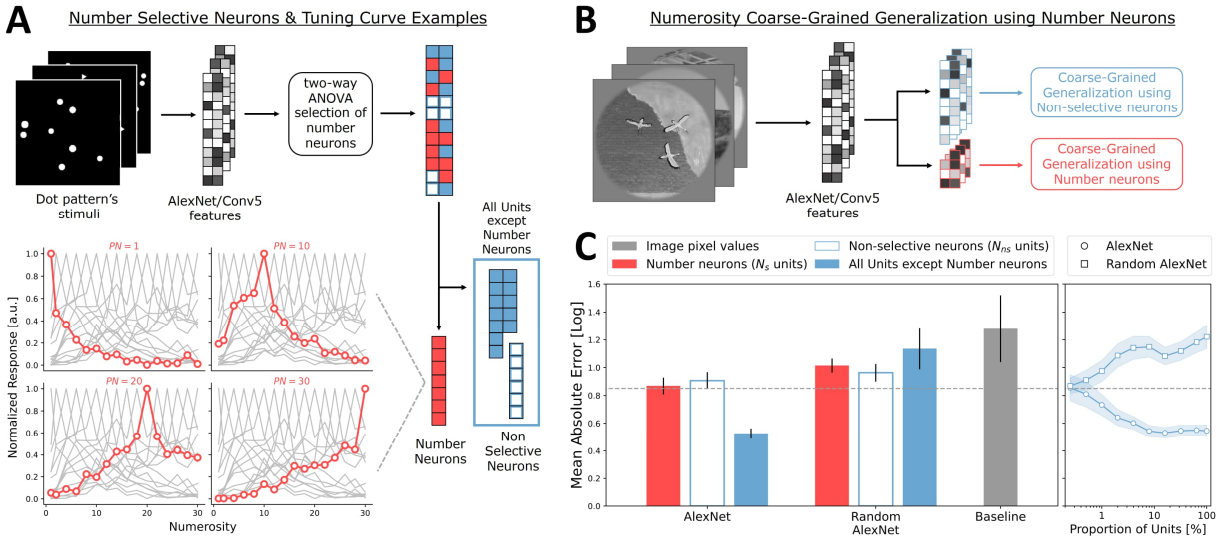

Fig. 4. Analyses of dot pattern selective units. (A) Analysis pipeline to identify number-selective neurons: Using visual dot arrays similar to previous studies, we identified artificial neurons that exhibited significant selectivity to numerosity in AlexNet's Conv5 (shown in red). Examples of the associated tuning curves for some preferred numerosities (PN) are also presented. Tuning curves are obtained by averaging the normalized activations of number-selective neurons which share the same preferred numerosity. **(B)** Overview of the coarse-grained generalization using subsets of neurons: The set of AlexNet/Conv5 artificial neurons is split into N_s number-selective neurons (shown in red) and the N_{ns} non-selective neurons (shown in blue). Ridge regression decoders are trained and tested to predict the number of objects across photorealistic stimuli using each subset of neurons. **(C)** Numerosity prediction performance (MAE) for the fifth layer (Conv5) of AlexNet and its untrained counterpart (Random AlexNet) for the coarse-grained generalization (means and STD over all four train-test combinations). Perfect performance is 0 and chance performance is indicated by a dashed line. MAE scores for decoding based on the N_s number-selective neurons and the N_{ns} non-selective neurons are shown in red and hollow blue, respectively. Decoding performance for the larger population excluding the N_s number neurons is shown in solid blue, and for the baseline of image pixel values in gray. MAE scores when randomly sampling different fractions of the entire artificial unit population are reported in the rightmost panel (means and std across all bootstrap iterations).

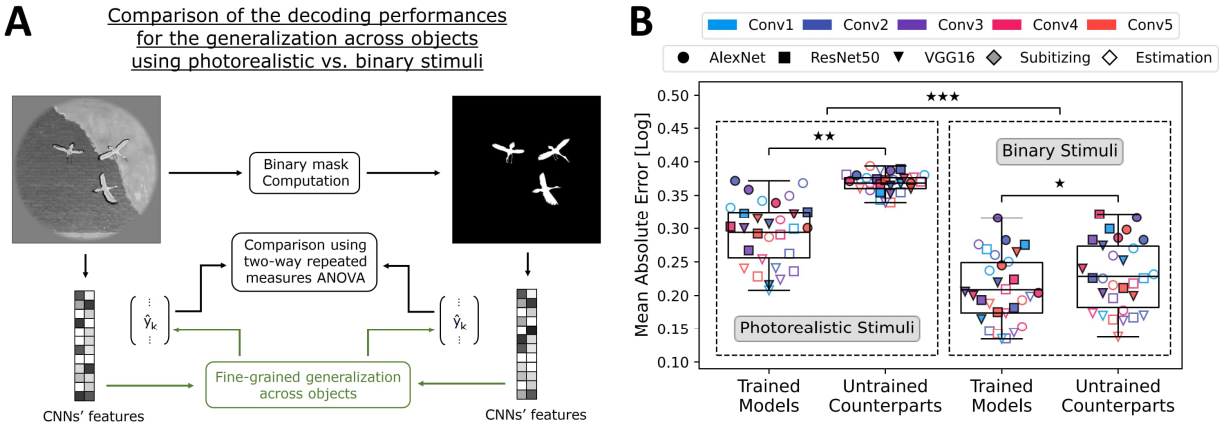

Fig. 5. Impact of using photorealistic compared to simplified stimuli. (A) Overview of the analysis: A binarized version is computed from each original stimulus by setting pixels within objects to white and the background to black. Following this, the feature representation for this simplified stimulus dataset is extracted from all CNN models. **(B)** Comparison of the numerosity decoding performance (MAE) of the different CNN architectures (markers) and layers depths (colors) for the fine-grained generalization across objects, using the photorealistic or the simplified stimulus datasets. Filled and hollow markers represent decoding performance when using the stimuli from the subitizing and estimation range, respectively. The boxes extend from the first quartile to the third quartile of the data, with a line at the mean. Significance for the difference between decoding on photorealistic and simplified stimuli corresponds to the interaction in a two-way repeated measures ANOVA: ***$p < 10^{-15}$. Significance for differences between trained and untrained models correspond to Wilcoxon signed-rank tests: *$p < 0.05$, **$p < 10^{-5}$.

Supplementary Materials for

Trained deep neural network models of the ventral visual pathway encode numerosity with robustness to object and scene identity

Thomas Chapalain *et al*.

Corresponding author: thomas.chapalain@inria.fr, evelyn.eger@gmail.com

This PDF file includes:

Figs. S1 to S4
Tables S1 to S2

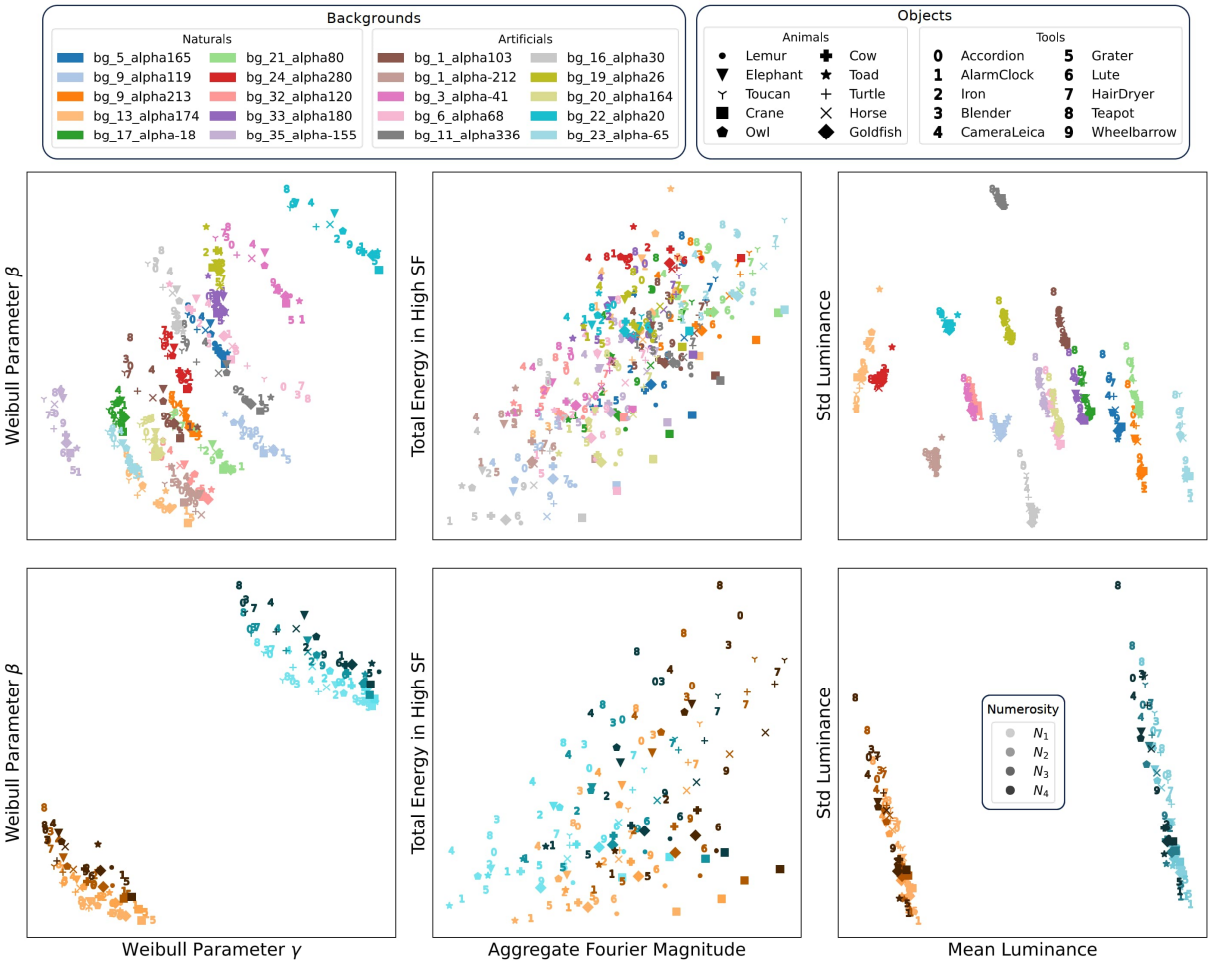

Fig. S1. Visualization of the characteristics of the photorealistic stimulus dataset in the space of low-level summary statistics of the images - subitizing range: 1, 2, 3 or 4 objects.

(Top row) Overview of the low-level statistics for every one of the 400 (object, background) pairs, depicted by a specific pair of (marker, color). Each plotted point represents the position of a given condition (object, background) averaged across the remaining parametric space dimensions (N, SzA, Sp) and the two versions of each stimulus (which vary only by the random locations of the objects) accounting for a total of 4x4x4x2=128 stimuli averaged.

(Bottom row) Overview of the low-level statistics as a function of numerosity for two selected backgrounds. Each plotted point represents the position of a specific condition (numerosity, object) averaged across the remaining non-numerical dimensions (SzA, Sp) and the two versions of each stimulus, accounting for a total of 4x4x2=32 stimuli averaged.

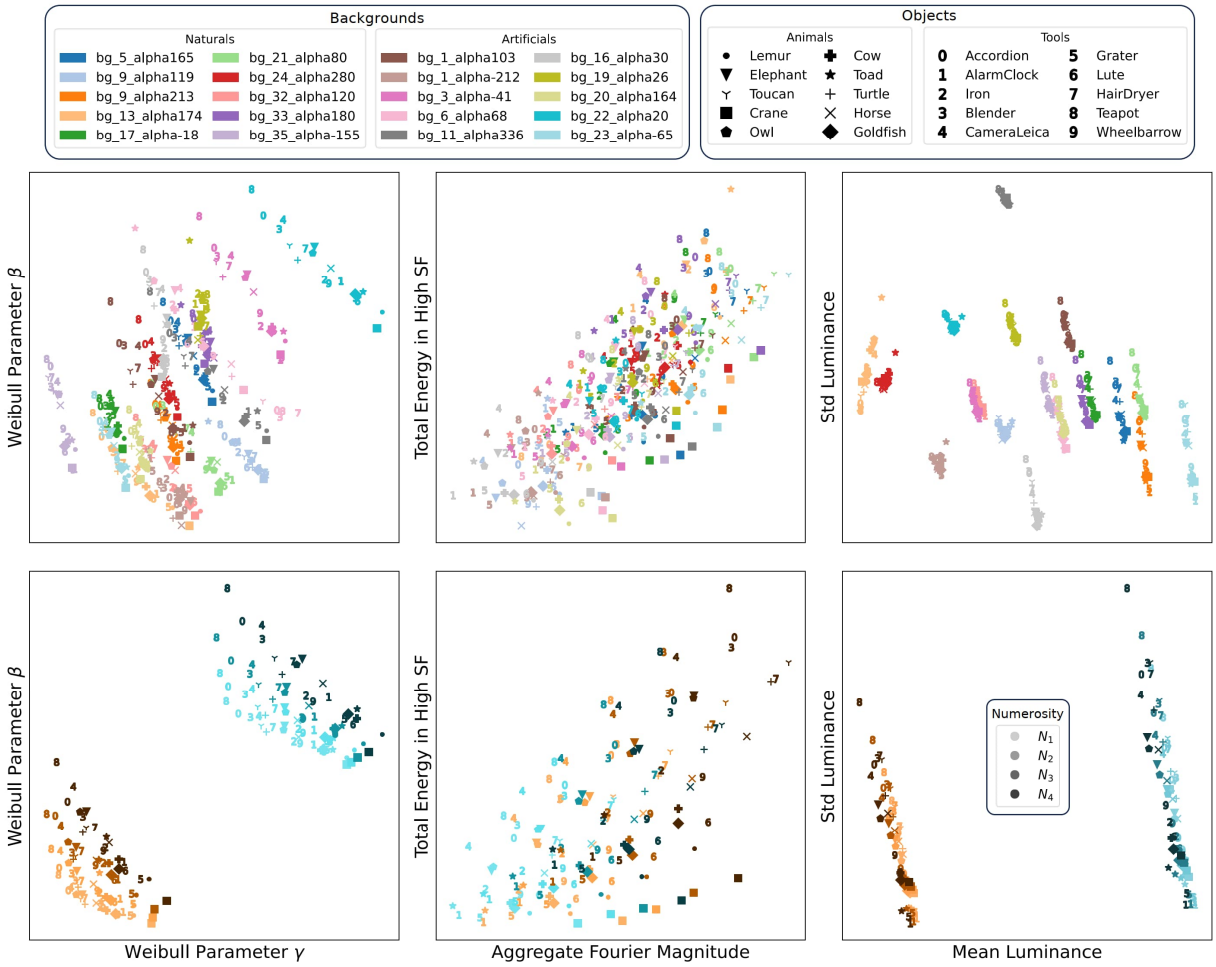

Fig. S2. Visualization of the characteristics of the photorealistic stimulus dataset in the space of low-level summary statistics of the images - estimation range: 6, 10, 15 or 24 objects.

(Top row) Overview of the low-level statistics for every one of the 400 (object, background) pairs, depicted by a specific pair of (marker, color). Each plotted point represents the position of a given condition (object, background) averaged across the remaining parametric space dimensions (N, SzA, Sp) and the two versions of each stimulus (which vary only by the random locations of the objects) accounting for a total of 4x4x4x2=128 stimuli averaged.
(Bottom row) Overview of the low-level statistics as a function of numerosity for two selected backgrounds. Each plotted point represents the position of a specific condition (numerosity, object) averaged across the remaining non-numerical dimension (SzA, Sp) and the two versions of each stimulus, accounting for a total of 4x4x2=32 stimuli averaged.

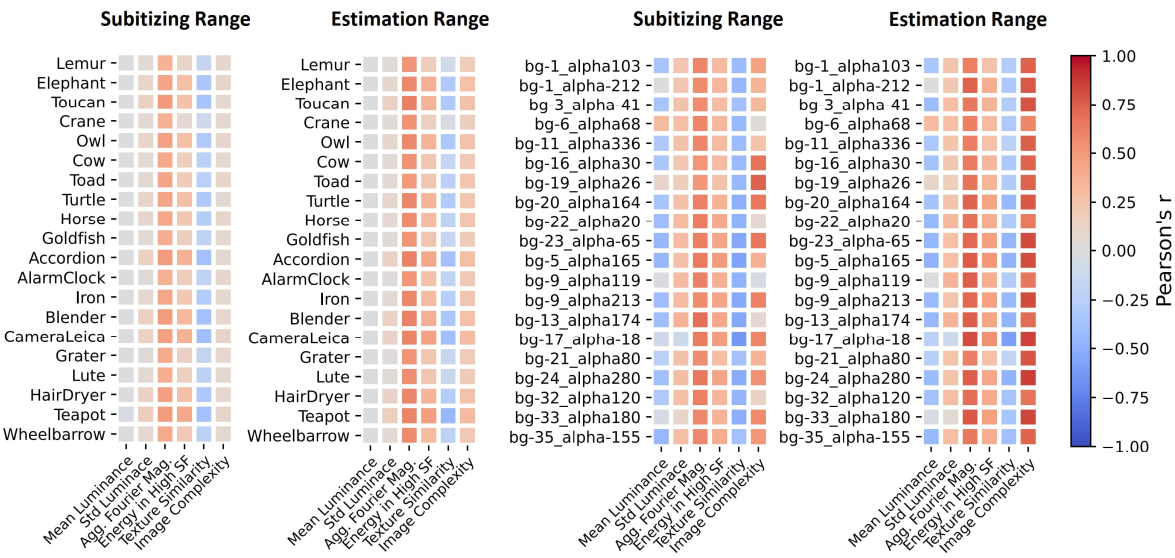

Fig. S3. Correlation between log-numerosity and low-level statistics per object/background

Summary of the Pearson correlation between log-numerosity and the six quantified low-level image statistics within our photorealistic stimulus datasets, computed for both numerosity ranges, subitizing (1, 2, 3 or 4 objects) and estimation (6, 10, 15 or 24 objects), across all stimuli sharing the same object (left two subplots) or background (right two subplots), respectively.

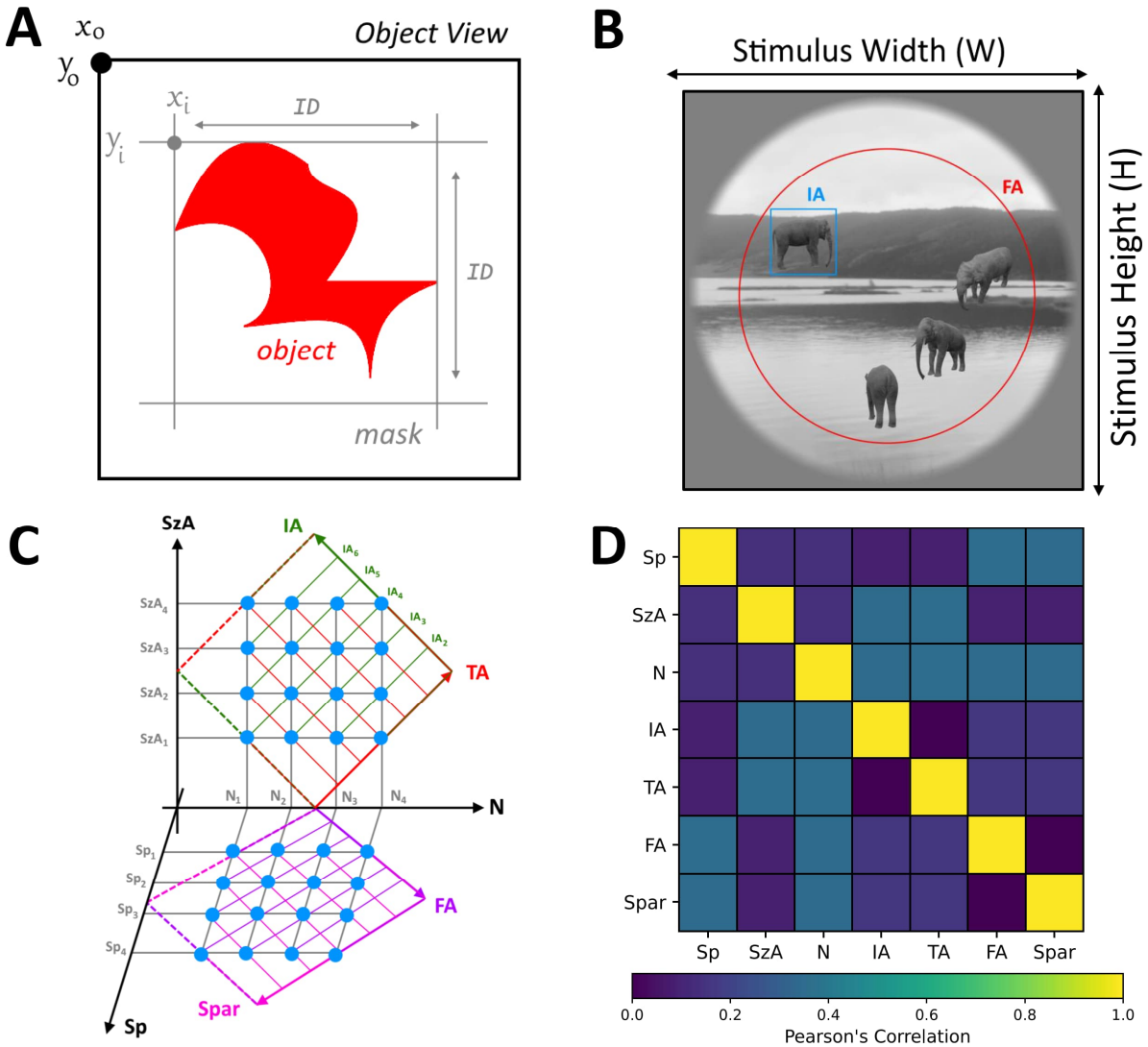

Fig. S4. Strategy to control non-numerical geometric stimulus properties in the stimulus dataset. Our stimulus design used a logic following Dewind et al., 2015, Cognition, for creation of multi-item patterns, adapted here to diverse object shapes rather than dots. **(A)** Definition of Item Diameter (ID): The 48 distinct views of a single object do not share an identical "physical" area, corresponding to the number of pixels encompassed by the object's shape (depicted in red). For the stimulus generation, we considered as the item diameter (ID) of an object the minimal size of the square (illustrated in grey) that encloses the shape of the object, and as item area (IA) the area of this square. **(B)** Stimulus example illustrating the field area (FA) corresponding to the circle within which the objects were allowed to fall, and the IA of one object. **(C)** The two dimensions "Spacing" (Sp, corresponding to an equal relative increase in field area FA and sparsity Spar) and "Size in Area" (SzA, corresponding to an equal relative increase in item area IA and total area TA) were varied orthogonally to numerosity N in four equidistant steps on a log scale. As a result, the constructs Sp and SzA are decorrelated from N whereas the individual non-numerical parameters (IA, TA, FA, Spar) are equi-correlated with N as can be seen in **(D)**.

Models	Extracted Layer Output	Layer Dimension	Convenience Name
AlexNet	1st Convolutional Layer	193600	Conv1
	2nd Convolutional Layer	139968	Conv2
	3rd Convolutional Layer	64896	Conv3
	4th Convolutional Layer	43264	Conv4
	5th Convolutional Layer	43264	Conv5
ResNet-50	1st Convolutional Layer	802816	Conv1
	1st Convolutional Residual Block	802816	Conv2
	2nd Convolutional Residual Block	401408	Conv3
	3rd Convolutional Residual Block	200704	Conv4
	4th Convolutional Residual Block	100352	Conv5
VGG16	1st Convolutional Layer	3211264	Conv1
	4th Convolutional Layer	1605632	Conv2
	7th Convolutional Layer	802816	Conv3
	10th Convolutional Layer	401408	Conv4
	13th Convolutional Layer	100352	Conv5

Table S1. Hierarchical convolutional networks and layers investigated. Detailed description of the layers and their respective dimensions used in our analyses for each one of the three CNN architectures: AlexNet, ResNet-50 and VGG16. All networks were used in the implementation available in PyTorch 2.0.

	Objects		Backgrounds	
Network	Test same	Test diff	Test same	Test diff
AlexNet trained	4 / 4	5 / 4	4 / 5	4 / 5
ResNet-50 trained	4 / 4	4 / 4	4 / 4	5 / 5
VGG16 trained	5 / 5	5 / 3	5 / 5	5 / 5
AlexNet untrained	3 / 5	5 / 5	5 / 5	4 / 2
ResNet-50 untrained	4 / 4	4 / 4	4 / 4	3 / 5
VGG16 untrained	5 / 5	2 / 2	5 / 5	5 / 5

Table S2. Most predictive layers in fine-grained generalization. Layers yielding the highest MAE score for each network when testing the decoder either one the same, or a different object/background (see Figure 3 of the main manuscript) are listed. The numbers separated by / refer to the convenience names of the convolutional layer as indicated in Table S1, for subitizing and estimation, respectively.