Do DNNs trained on Natural Images organize visual features into Gestalts?

Valerio Biscione and Jeffrey S. Bowers

Abstract  Gestalt psychologists have identified a range of conditions in which humans organize elements of a scene into a group or whole, and these perceptual grouping principles play an important role in scene perception and object identification. More recently, Deep Neural Networks (DNNs) trained on natural images have been proposed as compelling models of human vision based on reports that they perform well on various brain and behavioral benchmarks. Here we compared human and DNNs responses in discrimination judgments that assess a range of Gestalt organization principles. We found that most networks exhibited a moderate degree of Gestalt grouping for some complex stimuli at the last fully connected layer. However, in contrast with human neural data, this sensitivity vanishes at earlier visual processing layers. In a second experiment, by using simple dots configuration patterns, we found that all networks were only weakly sensitive to the grouping properties of proximity, and completely insensitive to orientation and linearity, three principles that have been shown to have a strong and robust effect on humans. Even top-performing models on the behavioral and brain benchmark Brain-Score miss these fundamental properties of human vision. Our overall conclusion is that, even when exhibiting Gestalt grouping, networks trained on 2D images use perceptual principles fundamentally different than humans.

Introduction  Human tends to group perceptual features together in order to form a coherent whole. Understanding when this happens has been the focus Gestalt psychology research for over 100 years and more than a hundred grouping “laws” have been suggested (Wagemans et al., 2012a). Whereas in the past the formulation of these laws were based on subjective experience and were criticised for a lack of scientific rigour, subsequent researchers have developed experimental designs with carefully constructed stimuli (e.g. Gabor displays, dot lattices) that allow for parametric control, richer visual displays, and objective measures of grouping effects (Wagemans et al., 2012b). One such approach consists of measuring the impact of salient Emergent Features (EFs) on discriminating visual patterns. These EFs derive from the relationship amongst individual parts rather than the parts themselves (Pomerantz et al., 1977; Pomerantz and Portillo, 2011). We will use the concept of EFs as the basis of our approach, as detailed later.

Recently there has been an explosion of interest in Deep Neural Networks (DNNs) as models of the human visual system for object recognition. Even though DNNs have primarily been designed to solve engineering tasks, reports that the pattern of activations of units in DNNs are similar to neural activation in human and macaque visual systems have led to the view that DNNs can be used as a test bed for simulating biological vision in mammals (Gauthier and Tarr, 2016; Kriegeskorte, 2015). As a way of formalizing this similarity, a Brain-Score benchmark has been put forward...
which has been enthusiastically embraced by researchers comparing DNNs to human vision. In the current paper, we explore several DNNs thought to be amongst the best model of human vision, and test whether they support various Gestalt grouping phenomena. In particular, we tested whether DNNs are sensitive to some basic principles of organization such as proximity, orientation, and linearity, and compare their responses to human responses from classic visual perception work (Pomerantz et al., 1977; Pomerantz and Portillo, 2011). We aim to tackle three interrelated questions:

1. Do DNNs exhibit human-like Gestalt grouping effects? Are they sensitive to the emergent properties of proximity, orientation and linearity?
2. Do models with a higher Brain-Score show a greater agreement with human similarity judgments in a Gestalt grouping task?
3. Do Gestalt grouping principles emerge as a consequence of learning statistical regularities of 2D natural images?

In the following sections we contextualize each of these questions and explain how our experiments address them.

**Neural Networks as a Model of the Human Visual System**

DNNs trained on ImageNet (a dataset consisting of 1000 categories of objects taken across over 1 million photographs, Krizhevsky et al. 2012) develop a set of internal feature representations that are statistically similar to the neural representations in human and non-human primate visual systems (Yamins and DiCarlo, 2016; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2018). A neuronal and behavioural benchmark called Brain-Score has been developed to assess neural networks on their similarity with biological object recognition systems, with DNNs performing much better than all previous approaches (Schrimpf et al., 2018). At the time of writing more than 160 models have been tested on Brain-Score.

In spite of these successes when tested on well known psychological phenomena, DNNs often fail on the most basic perceptual properties exhibited by humans. For example, DNNs do not possess human-like shape bias (Geirhos et al., 2018; Malhotra et al., 2020), they appear to discriminate categories based on local instead of global features (Baker et al., 2018b; Malhotra et al., 2022), do not account for humans' similarity judgments of 3D shapes (German and Jacobs, 2020), and fail to support basic visual reasoning such as classifying images as the same or different (Puebla and Bowers, 2021). In addition, DNNs often act in surprising non-human-like ways, such as being fooled by adversarial images (Szegedy et al., 2013; Dujmović et al., 2020) and make bizarre classification errors to familiar objects in unusual poses (Kauderer-Abrams, 2017; Gong et al., 2014; Chen et al., 2017). Furthermore, recent work by Xu and Vaziri-Pashkam (2021) failed to find strong neural correlates between DNNs' internal representation and fMRI signals from high-level visual areas of human participants.

At the same time, some studies have reported that DNNs capture some key psychological findings. For example Jacob et al. (2021) found that a DNN (VGG16) exhibited several important visual phenomena including the Weber's Law according to which the just noticeable difference between two varying stimuli is a constant ratio of the original stimulus. Biscione and Bowers (2021, 2022), found that networks exhibited strong invariance to several novel object trasformations (rotation, scale, change in luminance, translation, and to a lesser degree change in viewpoint), but only after being trained on a correspondingly transformed dataset of different classes, indicating that DNNs can learn the human perceptual property of object invariance to transformation (Blything et al., 2020, 2021). There are also reports of classic DNNs supporting an illusory motion phenomenon (Watanabe et al., 2018) and illusory contours (Kim et al., 2021; Baker et al., 2018a), as discussed next.
Neural Networks and Gestalt
As far as we are aware, Gestalt properties have been explored only in relation with illusory contours. Baker et al. (2018a) tested the degree in which networks could perceive illusory contours after being trained on non-illusory similar shapes (fat and thin rectangles). The network successfully predicted the type of shape regardless of whether the contour was normal or illusory, but the authors found no evidence that the network used the same information as humans, and concluded that CNNs do not perceive illusory contours. Kim et al. (2021) challenged this conclusion, and found that several architectures (including Inception Net) pretrained on ImageNet exhibited closure on display of edge fragments. Whether the later findings reflect Gestalt processes similar to human visual processing remains unclear (Lotter et al., 2020). Other researchers have focused on modifying architecture or training regime in order to explore this issue: Lotter et al. (2020) found that a network based on predictive coding (PredNet) which was trained on predicting the next frame of video sequences, exhibited disparate phenomena observed in the visual cortex, including the flash-lag effect and illusory contours. The same network also appeared to perceive illusory motion (Watanabe et al., 2018). Using a DNN modified to include feedback connections through predictive coding dynamics, Pang et al. (2021) also showed human-like perception of illusory contours. Illusory contours constitute an important phenomenon in Gestalt psychology, but human perception is mediated by a wider set of grouping principles which, to the best of our knowledge, have not been explored in DNNs.

If DNNs are going to be used as models of human vision it will be important that they not only do well on various brain-score measures but also account for key experimental results reported in psychology. Here we have focused on Gestalt rules of organization not only because they play a key role in visual perception and object recognition (Biederman, 1987; Perrett and Oram, 1993; Spillmann, 2009), but because there are existing image datasets and robust empirical phenomena that make it easy to test. Specifically, we consider whether DNNs show sensitivity to EFs, as described next.

Formation of Wholes through Emergent Features
Gestalt researchers have long studied the emergence of “wholes” from the combination of individual parts, but Gestalts have proven difficult to define and measure. Pomerantz et al. (1977) operationally defined Gestalts as the result of salient Emergent Features (EFs), that is features that are the result of the relations amongst individual elements, and are not possessed by the elements themselves. These EFs behave as though they were elementary themselves and are sometimes detected more quickly than the more basic features from which they arise. The effect can be measured in a discrimination task.

In Figure 1, left, we illustrate this idea. As a baseline, we test how well humans distinguish two stimuli (A and B, base pair). We then add a new contextual stimulus C to both the A and B images, creating a composite pair. Importantly, the stimulus C is not informative by itself in distinguishing AC from BC. Nevertheless, the composite pair is now much more discriminable than the base pair due to the interaction of the context with the base elements. Whenever the context stimulus changes performance on the discrimination task a Configural Effect (CE) is observed. Improved performance is described as a Configural Superiority Effects (CSE) and this provides a measure of an EF that is the product of a Gestalt organizational principle. By contrast, impaired performance is described as a Configural Inferiority Effect (CIE), and does not reflect an EF, but rather, reflects a number of possible factors, including additional computational and attentional load, increase similarity, crowding. That is, for a CSE to manifest itself, the EFs needs to be powerful enough to over-ride all these other factors. In a standard procedure, these stimulus sets have been used in a “oddity reaction time task” to measure CSE and CIE in various contextual configurations. In this experimental design, the subjects were asked to determine in which quadrant of a 2x2 grid an ”odd” stimulus was presented (Figure 1, right).

The advantage of this approach is that it provides a quantitative measures of CSEs/CIEs rather
Figure 1. Pomerantz et al. (1977) approach to measure Emergent Features (EFs). **Left**: starting with a base pair, a non-informative context is added to obtain a composite pair. **Right**: The base pair and composite pair are arranged in a 2x2 grid to form an odd-discrimination task. Participants are asked to indicate the location of the "odd" element, and RTs are measured for the base and the composite grid. With some composite grids (such as the one illustrated in the figure) discrimination is much faster than the corresponding base grid. Since the added context in the composite is equal for odd and non odd elements, any facilitation must be due to EFs.

than relying on subjective judgements. As predicted, many configurations (combination of characters, line segments forming letters, surfaces, 3D volumes) result in CIEs or very modest CSEs (Pomerantz and Portillo, 2011). However, critically, other configurations show strong CSEs (Pomerantz et al., 1977; Pomerantz and Pristach, 1989; Pomerantz and Portillo, 2011). Many of the strong CSEs observed with the complex stimuli used in Pomerantz et al. (1977) and here tested in Experiment 1 may be the result of a combination of multiple EFs and other factors, including high-level features such as shape familiarity or closure. In Experiment 2 we explore specific and low-level EFs (proximity, orientation, and linearity) with simple dot configuration following Pomerantz and Portillo (2011).

**Where do the laws of perception come from?**

A related question is the role of perceptual experience in acquiring EFs, that is, to what degree the grouping principles are learnt from the visual environment and to what degree they are a function of the innate architecture of the visual system (Todorović, 2011). Classic Gestalt psychologists minimized the significance of a learning account (Metzger, 1966). These authors conceded that some aspect of visual organization could be based on habit or learning (such as the ability to group particular continuous lines on a paper in letters), but these cases were thought to be the exception and weaker than others (Wertheimer, 1923).

Nevertheless, some evidence has emerged supporting the idea that basic Gestalt principles could be the result of applying statistical regularities acquired in everyday life: Peterson and Gibson (1994) found that a silhouette is more likely to be assigned as the figure if it suggests a common object (see also Peterson 2019); with two different paradigm, Duncan (1984) and Zemel et al. (2002) found that a perceptual grouping can be altered with only a small amount of experience in a novel stimulus environment. Other evidence based on RT responses have been collected by Vecera and Farah (1997). However, some other combinations are impenetrable to learning, as demonstrated by the Kanizsa stratification images (Grossberg and Zajac, 2017).

In the current work, all networks tested were pre-trained on ImageNet, a dataset consisting of thousand categories of objects taken across over 1 million photographs. Most of the networks used achieve an impressive degree of accuracy on a test set, at par with human performance on ImageNet (He et al., 2015). Therefore, regardless of their plausibility as a models of the human ventral pathways, we can use them to test whether learning statistical regularities on a complex domain of 2D natural images affords the network to extract EFs. Some of these regularities might be low-level such as being sensitive to proximity of two stimuli, or their orientations with respect to one another; other might be higher level such as grouping based on shape familiarity. If we fail to observe grouping principles in our experiments it could be that the models need to be trained on more complex datasets (e.g. an interactive 3D world) in order for some Gestalt principles to emerge, or alternatively, different DNN architectures are required.
Outline of the current work

We test a wide variety of DNNs on several sets of stimuli. Each set is composed of two pairs, a base and a composite (obtained by adding a non-informative context to the base pair as in Figure 1). We computed the cosine similarity of networks’ internal representation across each pair, to deduce the discriminability of the two images in the pair. By comparing the discriminability across the two pairs we obtained a Network Configural Effects (CE), that is the effect of adding the non-informative context to the base pair. This could result in either enhancing discrimination (CSE) or diminishing it (CIE). These measures can then be compared with CSEs/CIEs found in humans participants, assessed through reaction times (RTs) recording (more details in Methods). While we compared humans and networks on both CIEs and CSEs, notice that only CSEs are the result of Gestalt grouping, while CIEs correspond to crowding/attention load, etc.

We selected our models based on past claims that they are plausible models of the human visual system, either because the models appeared to support Gestalt effects in previous work, or because they have achieved a high score on the Brain-Score benchmark suite. The human RTs data were collected from two sources: Pomerantz et al. (1977) and Pomerantz and Portillo (2011). For Experiment 1, we used 17 Sets of stimuli from Pomerantz et al. (1977) each composed of a base and a composite pair of images (Figure 3, left). Five of these sets generated high CSEs in humans, indicating a strong Gestalt grouping effect. In Experiment 2, we generated a wide number of configurations composed of simple dot patterns, as introduced by Pomerantz and Portillo (2011) and illustrated in Figure 5 and in Methods). The structure of each set allowed for the investigation of specific emergent features, excluding confounding effects due to the complex nature of stimuli used in Experiment 1. Pomerantz and Portillo (2011) found a strong and consistent effect for three EFs: proximity, orientation, and linearity, and therefore we tested whether these same features could be used by the networks to enhance discriminability of a base image pairs. We generalized the results across different background conditions and transformation conditions (rotation, translation, scale, and no transformation).

Results

Experiment 1

CSEs are a measure of Gestalt grouping, and so we focused on the 5 sets that exhibited the strongest CSE in humans (set 1 to 5 in Figure 3, left). In humans, these set produced CSEs from 0.7 to 1.38 seconds, corresponding to a speed up between 40% to 180%). To match human perception, all these sets should elicit large CSEs in the networks. When analysing the last fully connected layer, sets 1, 2 and 3 indeed produced large CSEs for most network (the clear exception being AlexNet), with the strongest effect shown by InceptionNet for sets 1 and 2. However, sets 4 and 5 produced either no effect or slightly inferiority effect (Figure 2, top). We also computed the same analysis for an earlier stage, the last convolutional layer, finding that the CSEs for Sets 1 and 2 disappeared for most networks (with the exception of Inception Net), becoming CIEs instead (Figure 2, bottom). At earlier layers the effect completely vanishes for all networks (see Appendix). We extended the analysis to the full range of stimulus sets by plotting humans CEs vs the networks CEs in Figure 3 (right) for the last layer (a fully connected layer, circles), and the last convolutional layer (triangles). Consistent with humans, networks did not exhibit strong Gestalt grouping with any other Set, but the amount of CIE was inconsistently matching human responses. Interestingly, the set producing the highest CIE in humans (Set 17) produced a moderate degree of CSE for some networks.

To quantify whether humans and networks CEs are correlated, we computed the Spearman’s rank correlation coefficient across human and networks’ CEs. Overall, apart from AlexNet, the relationship is positive, (~ 0.55, indicating a moderate agreement) but non-significant at p < 0.01 (only exception being InceptionNet with p = 0.007).

We repeated the same analysis at earlier convolutional network layers (at three equidistant stages preceding the last convolutional layer). The results matched the outcome of the last convo-
lutional layer presented here: most of the stimulus sets that produced CSEs in humans resulted in CIEs in the networks. No model showed strong CSEs at any layer earlier than the last fully connected layer (see Figure 7, the only exception being Inception Net, and only for the very last convolutional layer). This is in contrast with human neural data, for which Gestalt grouping appears to be an early process (Alexander and Van Leeuwen 2010; Fox et al. 2017). The robustness of these results is confirmed by identical outcomes across different background and transformation conditions.

Overall, this Experiment produced mixed results. Apart from AlexNet a similar pattern of CEs was obtained across all networks, with only a subset of the stimulus sets that produce large CSEs in humans evoking high CSEs in networks, and only at the final layer. Instead, the most common outcome was CIE in DNNs, and the pattern here sometimes matched human CIEs, and other times not. And in any case, CIE do not assess EFs/Gestalt grouping.

How can the CSEs for Stimulus Sets 1-3 in the final layer be explained? One hypothesis is that, by being trained on ImageNet, networks acquired a strong sensitivity to some basic shapes, making these shapes salient and discriminable when compared to different shapes or non-shapes. This can explain the large CSE for Set 1 and Set 2 (a triangle versus a non-shape), and Set 3 could possibly be perceived as an ellipse (especially considering that DNNs appeared to exhibit Gestalt closure, see Kim et al. 2021). This hypothesis is particularly intriguing as it implies that sensitivity to complex, abstract shapes (Weisstein and Harris, 1974) may arise simply by learning the statistics on 2D images when performing classification.

We tested this hypothesis by modifying Sets 1, 2 and 3 so that they did not contain familiar shapes. The results (Figure 4) for Set 2 and 3 (Set 1 results are similar to Set 2 results and thus omitted) indicates that most networks still exhibited Gestalt grouping in spite of objects not resembling abstract familiar shapes (although with a higher variability for Set 3). Similarly to the
original stimulus sets, no network but Inception Net exhibited any grouping at the convolutional
layer for Sets 1 and 2, and most networks did for Set 3. These results suggest that the CSEs ob-
served in the final layer do not reflect sensitivity to abstract shape, but might be related to other
effects such as sensitivity to closed shapes, to sharp angles, curved lines, etc.

Overall, the above pattern of results suggests important differences between perceptual group-
ing in humans and networks. The lack of a CSE for stimulus Sets 4 and 5 indicates that humans
perceive EFs that networks are insensitive to. At the same time, it is unclear whether the CSEs
observed with Sets 1, 2 and 3 are due to the same EFs employed by humans given the complex-
ity of the stimuli. The finding that these effects are only observed in later layers of the networks
highlights that they are playing a different function in classifying objects compared to the role of
Gestalts in perceiving and classifying objects in humans (Biederman, 1987).

The main limitation with the first experiment is that the stimuli are quite complex, and accord-
ingly, the CSE could reflect a complex interaction of multiple EFs as well as other factors. In order

---

**Figure 3.** Left: the full set of base and composite pairs used in the experiment to assess Configural Effects (CE) in networks. Sets are sorted by the amount of CEs elicited in humans, measured in Pomerantz et al. (1977). Gestalt grouping is measured by CSE. Right: networks and humans' CE for the 17 sets, at the last convolutional layer and the last layer (black-on-random-pixel background, translation condition). Points on the top-right area indicates human and network agreements on Gestalt grouping; points on the bottom-left area indicates agreement on crowding/ context interference. The Spearman's rank-correlation coefficient $r_s$ for the last layer indicates a moderate agreement at this stage (orange circles), but this is almost totally driven by 5 specific sets (see text). Gestalt grouping agreement is almost absent at an earlier stage (blue triangles).
to assess exactly whether humans and networks are sensitive to the same grouping principles, we designed Experiment 2 in which individual EFs are isolated.

**Experiment 2**

In order to explore to study EFs more selectively, Pomerantz and Portillo (2011) designed an odd-discrimination task in which dot patterns were used to create base and composite pairs. They found that CSEs were consistently exhibited for three EFs. Specifically, when compared to the base pairs, orientation, proximity and linearity respectively generated a CSE of 0.16, 0.41 and 0.36 seconds, corresponding to a speed up of 11%, 29% and 26%, respectively (the smaller CSEs here compared to the previous stimuli are not surprising given that they were based on isolated EFs). Stimuli generation is illustrated in Figure 5 and in more details in the Methods. The same three background conditions from the previous experiment were used.

**Preliminary Tests**

We performed two preliminary tests. We firstly assessed whether the dot stimuli employed in this experiment were salient enough to be discriminated by the networks. Indeed, the cosine similarity between a pair with an empty canvas and a canvas with a single dot was much lower than the cosine similarity between a pair of empty canvasses (with pixellated background), confirming the saliency of the dot stimulus (see Figure 8). Furthermore, we verified that the tested EFs of proximity, orientation and linearity can be learnt when directly training each network to distinguish them (see Methods). We found that almost all networks (with the exception of AlexNet) quickly learnt to classify based on proximity, orientation or linearity, reaching an accuracy of around 99% on the test set.

**Results**

Pomerantz and Portillo (2011) observed robust CSEs in humans for all conditions in their behavioral studies: that is, adding a second dot to generate proximity or orientation features resulted in faster RTs (higher discriminability) compared to the base pair, and adding a third dot to induce a linearity feature resulted in even faster RTs. We computed the Network CEs by comparing each pair with
Figure 5. Generation of stimuli for Experiment 2, following Pomerantz and Portillo (2011). Starting with a pair of images in which the only discriminant feature is the location, an additional dot is added, yielding the EF of proximity or orientation. The EF of linearity is obtained by adding a dot to the orientation pair. These three EFs have been found to elicit strong and consistent grouping effects in humans (Pomerantz and Portillo, 2011). Each network was tested on 100 randomly generated sets, and each analysis was repeated on three different stroke-over-background conditions (see text).

the base one (single dots). For a network to match human behaviour, it should produce high CSEs across all three EFs conditions.

Figure 6. Amount of Configural Effect (CE) in Networks. In humans, all three of these conditions produced high CSE. In networks, the CSE was significant only for proximity, with an extremely small effect size, and only in the last fully connected layer (top). In earlier layers, the context hindered discrimination rather than facilitating it (bottom, see also Figure 9 in the Appendix.

The results are shown in Figure 6 for the last (fully connected) layer (top) and the last convolutional layer (bottom). While proximity seemed to produce significant effects for some networks, notice that the effect size is extremely small. This is not the result of a ceiling effect (cosine similarity is bound from -1 to 1): if proximity made the composite pair more discriminable, the cosine similarity value should decrease compared to the value of the base pair. Given that the cosine similarity of the base pair is equal to 0.9, we can be sure that there is plenty of space for this metric to decrease, and the small effect size was not an artifact of the metric used.

1 This high value is explained by considering that the base pair contains two images that differ only on dot location. The networks tested here were pretrained on ImageNet which was shown by Biscione and Bowers 2021 to provide the networks with translation tolerance, a property also possessed by humans (Blything et al., 2021). This produced similar internal representation across different dot locations, and thus high cosine similarity.
Across all other layers, the models appeared to be sensitive to additional features in the opposite way as humans (see Figure 6, bottom, for the last convolutional layer, but similar results were observed across the whole networks’ depth as shown in Appendix). That is, adding one or two dots to a canvas with a single dot, inducing either proximity, orientation, or linearity, does not help discrimination but hinders it. Furthermore, the same trend was observed regardless of the background condition.

Discussion
We will now elaborate to what extent our experiments can answer the questions raised in the Introduction and then discuss how our findings relate to previous work assessing Gestalt organizational principles in DNNs.

Do modern DNNs exhibit Gestalt grouping principles?
Our key finding is that DNNs failed to support Gestalt grouping for many stimulus sets that produce Gestalts in humans, and for the subset of stimuli did generate Gestalts in DNNs, the processes that gave rise to them and their function appear quite different than in humans. In particular, most DNNs trained on ImageNet exhibited a Configural Superiority Effect (CSE) for three out of the five stimuli to which humans are particularly sensitive to (Experiment 1). However, this was only observed at the last fully connected layer, and the effect vanishes at early processing stages. The complex stimuli used in this Experiment were not suited for understanding what feature drove the perceptual grouping. In Experiment 2, by using simpler dot configuration patterns, we found that networks were very mildly sensitive to proximity (again only at the last fully connected layer), and not sensitive to the Emergent Features (EFs) orientation and linearity at any layer.

It is important to emphasize that EFs are powerful Gestalt grouping phenomena that are not only subjectively compelling (see Figure 1) but that support fast RTs in participants, similar to other low-level visual features that “pop out” (Treisman, 1998). These EF (and Gestalt rules more generally) are generally though to be computed in early level vision (Alexander and Van Leeuwen, 2010; Fox et al., 2017) and are thought to support the figure-ground segregation (Wagemans et al., 2012a) as well as building representations of object parts (Biederman, 1987). Accordingly, the finding that DNNs only showed EFs for a subset of stimuli, and only at later layers, reflects a fundamental difference between how current DNNs model and human visual system. Note, Inception Net did support some EFs in both the fully connected layer and the last convolutional layer. However, even in this case none of the earlier layers exhibited Gestalt grouping, as shown in Figure 7, and similarly to other networks, it exhibited very limited to no sensitivity to EFs in Experiment 2.

Is Brain-Score a good predictor of whether a DNN supports EF?
We tested a range of DNNs, some of which performed well on the Brain-Score competition designed to assess DNN-human correspondence in object recognition. However, there is no sign that better Brain-Scores were associated with larger EFs. Rather, what emerged from both Experiment 1 and Experiment 2 was a general agreement amongst networks: most networks showed Gestalt grouping for the same small subset of stimuli Experiment 1, and only for the last fully connected layer (with the exception of InceptionNet, which in fact ranks lower in Brain-Score than most other networks tested here). Similarly, most networks showed very mild or non-existent effect on proximity, and null or negative effect of orientation and linearity in Experiment 2, with no relation to Brain-Score. This suggests that the Brain-Score benchmarks are missing important psychological tests that measure fundamental features of human vision. Indeed, it could be argued that accounting for these phenomena is of higher relevance than observing neural correlates or similarity in classification accuracy. This adds to a long list of other important psychological findings that the DNNs with the highest brain-scores fail to explain, as briefly reviewed in the Introduction.
Acquisition of Gestalt Properties
The degree to which some basic grouping properties can be learnt has been a controversial topic for many decades (see Introduction). Whether or not DNNs are good models of the human visual system, they are excellent at extracting statistical regularities from a dataset, and therefore provide a test to whether some grouping phenomena are implicitly encoded within the statistics of a particular dataset.

We found that networks trained on ImageNet developed a sensitivity to some particular stimuli configuration, (Experiment 1), regardless of the background they were presented in, and across several transformations. Since shape familiarity is a principle used by humans in perceptual grouping (Weisstein and Harris, 1974), we verified whether the observed grouping was the result of acquiring sensitivity to some abstract shapes through being trained on natural images. By slightly changing the stimuli in order to make them unfamiliar, we found this not to be the case. But the critical point for our purposes is that training on ImageNet only induced very weak EFs for proximity, and none for orientation and linearity that are so salient for humans.

The fact that the networks appear to show some Gestalt grouping for some stimuli and not others suggests that training on natural 2D images might induce different grouping properties than those experiences by humans. Still, it is possible that using a more realistic dataset (e.g. a 3D environment), or more naturalistic training environments (e.g. with a reward signal such as in the reinforcement learning approach) could result in the acquisition of a wide array of grouping principles, but this is pure speculation for now.

In addition, the finding that the current DNNs did not learn the current set of EFs does not imply that other properties such as closure could not be, with the same 2D dataset. Gestalt laws of organisation are not acquired at the same time in humans, but appear at different developmental stages. For example, in humans, 3-4 months old infants can use the principle of proximity (Hayden et al., 2008), but not good continuation, good form or good similarity (see also Quinn and Bhatt (2005)). Whether these different developmental stages implicate learning or the maturation of the visual system is unclear, but it highlights that only a subset of Gestalt laws may be learned in DNNs. The work Kim et al. (2021) seem to indicate that networks trained on ImageNet exhibit closure, and our findings lend some support to this idea given the CSE observed with unfamiliar but closed shapes. However, see Baker et al. (2018a) for a different account.

It is also interesting to notice that Gestalt grouping has been obtained in artificial networks that have a very different architectures than DNNs (Grossberg et al., 1997; Francis et al., 2017; Herzog et al., 2003). In other cases, ad-hoc architectural modifications have been added do DNNs to solve task that appear to underlie Gestalt grouping (Linsley et al., 2018). This suggests that Gestalt principles might be obtained only with an appropriate architecture, and cannot simply be extracted from the statistics of the visual environment. The current work suggests that standard DNNs with high Brain-Score might not fall within this category.

Conclusion
Overall, the results presented here highlight a striking disconnect between DNNs trained on natural 2D images and human human vision. In spite of testing DNNs that perform well on the behavioral and brain benchmark Brain-Score, our experiments show that these same DNNs fail to support a range of Gestalt grouping effects that play a fundamental role in human visual experience and object identification. Networks did show Gestalt-like responses to some stimuli (Experiment 1), but the pattern of responses suggests that these effects were the product of different grouping processing compared to human vision. Furthermore, when the emergent features of proximity, orientation, and linearity were individually investigated (Experiment 2), networks failed to show Gestalt effects, and in fact, the added context hindered discrimination at all but the last layer. Humans, by contrast, show robust Gestalt effects with these stimuli. This work highlight the importance of comparing networks performance with well established psychological phenomena, which have been largely ignored when comparing DNNs to the humans brain.
Methods

Code is provided in full\(^2\). In the code, stimuli are generated at runtime, and therefore no dataset is needed to replicate the results. However, a dataset of the stimuli used in the experiments is provided\(^3\).

Stimuli Generation

Experiment 1
We used the image pairs reconstructed from Experiment 1 and Experiment 2 of Pomerantz et al. (1977). Images were arranged in 17 sets, each composed of two pairs: a base and a composite pair (the full set is shown in Figure 3, left). The sets were composed so that they could elicit a wide variety of CSEs and CIEs. Each image was scaled so that its size would correspond to the size each network was trained on. Each analysis was repeated across three stroke-over-background conditions: white on black, black on white, and black on a random pixel background. Furthermore we used 4 different transformation conditions: no transformation, translation (18% of the image size), scale (0.7 to 1.3), and rotation (up to 360 degrees). The same transformation was applied to both images of each pair. For the conditions employing transformation, each pair was tested 100 times.

Experiment 2

Stimuli generation
Following Pomerantz and Portillo (2011) we generated pairs of stimuli starting with the simplest pair consisting of a single dot at different location. We added a context canvas that would elicit either the Emergent Feature of proximity, orientation, or linearity (Figure 5). We used three stroke-over-background conditions: white over black, black over white, and black over a background of random pixels, and generated a vast array of sequences. For each network, the cosine similarity was averaged across 100 sequences. Each dot in the base configuration was constrained to be located at a distance of at least 20 pixels from one another, and 40 pixels from the border in order to avoid border effects (Kayhan and van Gemert, 2020).

Network Used

We selected 8 networks based on their historical importance, their performance on standard dataset, and their biological plausibility. We used as a point of reference the Brain-Score value (Schrimpf et al., 2018), indicating the amount of variance explained by the model across several benchmarks. AlexNet (Krizhevsky et al., 2012), VGG19 (Simonyan and Zisserman, 2014), and ResNet (He et al., 2016) (we used ResNet-152) are classic networks that have been often tested on several cognitive phenomena (Schrimpf et al., 2018; Baker et al., 2018a; Biscione and Bowers, 2022), with mixed results. InceptionNet (Szegedy et al., 2015) was shown by Kim et al. (2021) to be sensitive to the effect of Gestalt Closure and thus it seemed suited for this battery of tests (we used InceptionNet V3). In DenseNet (Huang et al., 2017) each convolutional layer is connected with each other layer. A smaller version of this family of networks, which we used (Densenet-121), has been showed to possess human like translation invariance (Biscione and Bowers, 2021).

We also tested two models specifically developed to be biologically plausible and that provide a good match with primate neural data. The “CorNet” model family (Kubilius et al., 2019) aimed to incrementally build a network architecture by adding recurrent and skip connections while monitoring both classification accuracy and agreement with a body of primate brain neural data. From this family, the CorNet-S was selected as the best CorNet architecture.

The VOneNet family (Dapello et al., 2020) has been developed to better match the structure of the primate visual cortex. Each VOneNet contains a fixed weight neural network front-end that simulates primate V1, called the VOneBlock, followed by a neural network back-end adapted

---

\(^2\)https://github.com/ValerioB88/gestalt-DNNs

\(^3\)https://valeriobiscione.com/PomerantzDataset
from current CNN vision models. We used two versions of VOneNet: one with CorNet-S backend (VOneNet-Cornet-S), and the other with Resnet50 backend (VOneNet-Resnet50). VOneNet-Resnet50, DenseNet, ResNet-152 and VGG19 are in the top 10 on Brain-Score at the moment of writing, all with a score of \( \sim 0.45 \) (the highest scoring network obtained 0.468\(^4\)). InceptionNet V3, Cornet-S and its VOneNet scored slightly lower (\( \sim 0.42 \)). Finally, amongst the model we used, AlexNet scored the lowest (0.406). All networks were pretrained on ImageNet.

**Cosine Similarity and RT correlation**

We used a set of stimuli based on Pomerantz and Portillo (2011) and Pomerantz et al. (1977), and compared network performance with human RTs extracted from those works. In their work, the RTs were obtained by using a 2x2 grid as explained in the Introduction: the participants were asked to press a key on a keyboard indicating which quadrant contained the odd stimulus, and the RTs were recorded. In presenting these images to networks, we omitted the 2x2 grid and used pairs of images instead (stimuli generation is detailed above). The main difficulty in comparing Pomerantz's behavioural results with neural networks is that the behavioral results were based on RTs, which DNNs do not produce (there are some exception, but the models commonly scoring high on Brain-Score do not possess this feature). However, since we have direct access to models' internal representations, we can nevertheless obtain a measure of stimuli discriminability. We presented a pair of images to the network; for each image, we recorded the value of activation for every unit of a given layer, obtaining an activation vector for each image and each layer; The “distance” between the two activation vectors would correspond to a measure of discriminability for the image pair at a particular layer. We compared these measures to human RTs: high distance would correspond to high discriminability, which would correspond to fast RTs; and viceversa.

As a distance metric between internal representation we used the cosine similarity:

\[
C_l(a, b) = \frac{d_l(a) \cdot d_l(b)}{||d_l(a)|| \cdot ||d_l(b)||},
\]

where \(d_l(a)\) is the activation at layer \(l\) given input vector \(a\).

We then used the cosine similarity value to measure discrimination effect due to grouping as described below. When feeding the images into the networks, we first resized them to the same size used during networks' pretraining (that is 224x224 for all networks but InceptionNet, which was 299x299). Furthermore, all images were normalized with mean and standard deviation used during ImageNet pretraining.

**Comparing Configural Effects**

By using the same approach outlined in Pomerantz and Portillo (2011), we obtained the networks’ Configural Superiority Effects (CSEs) and Configural Inferiority Effects (CIEs) by computing the difference in cosine similarity across two pairs of stimuli: a base and a composite pair. The composite pair is obtained by adding a non-informative feature to each image of the base pair. If the composite pair similarity is more discriminable (lower similarity) than the base pair we would obtain a CSE, otherwise a CIE. We refer to the general difference across composite and base as Configural Effect (CE). For humans, CE is simply computed as \(RT_{base} - RT_{composite}\) with positive values indicating CSE and negative CIE. Since both high RTs and high cosine similarity correspond to lower discriminability, we measured the network CE as:

\[
NetworkCE = C_l(base, base) - C_l(composite, composite).
\]

\(^4\)We planned to test the top 5 ranking models on the Brain-Score benchmark at the time of writing. Amongst these, we could only test ResNet-152. The first two VOneNet models are not publicly available and it is not clear what the third place (ResNet-50-robust) networks refers to. We contacted the authors but they have not provided additional information. Nevertheless, 4 of the networks tested are in the top-ten at the time of writing.
Where each argument to \( C \) is an image for one of the pair (e.g. the left image of base pair in \( \text{Figure 1} \) would be \( \text{base}_a \), the right image \( \text{base}_b \), and \( C \) is the cosine similarity at layer \( l \). Therefore, high CSE in humans would correspond to high CSE in networks.

Agreement between networks’ and humans’ CE can be visualized graphically. Consider any of the subplots in \( \text{Figure 3} \), right. Any points falling on the upper-right or lower-left quadrant corresponds to networks and humans having the same type of CE (respectively CSE and CIE). Points falling on the upper-left or lower-right would indicate disagreement. Notice that Gestalt grouping is measured by high CSEs, whereas CIE are generally related to interference effects (crowding, attention load).

**Training on Proximity, Orientation, and Linearity**

Before running Experiment 2 we run a preliminary study, testing whether the EFs of proximity, orientation and linearity can be learnt in principle, by directly training each network to distinguish them. In fact, if a network were not able to distinguish different levels of e.g. proximity when directly trained on it, it will also not be able to implicitly extract this features when trained on a general classification task of natural images.

We used the same types of dots stimuli used in the rest of Experiment 2, which generation is described above. In the “proximity” condition we trained the networks on 6000 images divided in 3 classes: close (distance of less than 50 pixels between the two dots), medium (distance between 60 and 110 pixels), and large (more than 120 pixels). We similarly divided the “orientation” dataset in three classes (between 0 degrees and 25 degrees, between 35 degrees and 60 degrees, and between 75 degrees and 90 degrees). We trained on "linearity" by having two classes: either 3 dots linearly arranged or not linearly arranged. The networks were subsequently tested on 200 held out samples for each condition. We used Adam optimizer (learning rate=5e − 4) with Cross Entropy Loss, and training until convergence.

**Acknowledgement**

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 741134).

**References**

Alexander DM, Van Leeuwen C. Mapping of contextual modulation in the population response of primary visual cortex. Cognitive Neurodynamics. 2010; 4(1):1–24. [record/2010-02918-001], doi: 10.1007/S11571-009-9098-9.

Baker N, Erlikhman G, Kellman P, Lu H. Deep Convolutional Networks do not Perceive Illusory Contours. Cognitive Science. 2018;

Baker N, Lu H, Erlikhman G, Kellman PJ. Deep convolutional networks do not classify based on global object shape. PLoS Computational Biology. 2018; 14(12):1–43. doi: 10.1371/journal.pcbi.1006613.

Biederman I. Recognition-by-Components: A Theory of Human Image Understanding; 1987.

Biscione V, Bowers JS. Convolutional Neural Networks Are Not Invariant to Translation, but They Can Learn to Be. Journal of Machine Learning Research. 2021; 22(229):1–28. [http://jmlr.org/papers/v22/21-0019.html].

Biscione V, Bowers JS. Learning online visual invariances for novel objects via supervised and self-supervised training. Neural Networks. 2022 3; [http://arxiv.org/abs/2110.01476], doi: 10.1016/J.NEUNET.2022.02.017.

Blything R, Biscione V, Bowers J. A case for robust translation tolerance in humans and CNNs. A commentary on Han et al. arXiv preprint arXiv: 201205950. 2020 12; [http://arxiv.org/abs/2012.05950].

Blything R, Biscione V, Vankov II, Ludwig CJH, Bowers JS. The human visual system and CNNs can both support robust online translation tolerance following extreme displacements. Journal of Vision. 2021; 21(2):1–16. [https://doi.org/10.1167/jov.21.2.9.], doi: 10.1167/jov.21.2.9.
Chen FX, Roig G, Isik L, Boix X, Poggio T. Eccentricity dependent deep neural networks: Modeling invariance in human vision. AAAI Spring Symposium - Technical Report. 2017; SS-17-01 -.S41–S46.

Dapello J, Marques T, Schrimpf M, Geifer F, Cox DD, Dicarlo JJ. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. 34th Conference on Neural Information Processing Systems (NeurIPS 2020). 2020; https://github.com/dicarlolab/vonenet.

Dujmović M, Malhotra G, Bowers JS. What do adversarial images tell us about human vision? eLife. 2020 9; 9:1–29. doi: 10.7554/ELIFE.55978.

Duncan J. Selective attention and the organization of visual information. Journal of experimental psychology. General. 1984; 113(4):501–517. https://pubmed.ncbi.nlm.nih.gov/6240521/, doi: 10.1037//0096-3445.113.4.501.

Fox OM, Harel A, Bennett KB. How configural is the configural superiority effect? A neuroimaging investigation of emergent features in visual cortex. Frontiers in Psychology. 2017 1; 8(JAN):32. doi: 10.3389/FP-SYG.2017.00032/BIIBTEX.

Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. Psychological Review. 2017 7; 124(4):483–504. doi: 10.1037/REV0000070.

Gauthier I, Tarr MJ. Visual Object Recognition: Do We (Finally) Know More Now Than We Did? Annual review of vision science. 2016; 2(October):377–396. doi: 10.1146/annurev-vision-111815-114621.

Geirhos R, Medina Temme CR, Rauber J, Schütt HH, Bethge M, Wichmann FA, Temme CRM, Rauber J, Schütt HH, Bethge M, Wichmann FA. Generalisation in humans and deep neural networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. 31 Curran Associates, Inc.; 2018. p. 7538–7550.

German JS, Jacobs RA. Can machine learning account for human visual object shape similarity judgments? Vision Research. 2020; 167(December 2019):87–99. https://doi.org/10.1016/j.visres.2019.12.001.

Gong Y, Wang L, Guo R, Lazebnik S. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014 3; 8695 LNCS(PART 7):392–407. http://arxiv.org/abs/1403.1840.

Grossberg S, Mingolla E, Ross WD. Visual brain and visual perception: how does the cortex do perceptual grouping? Trends in neurosciences. 1997 3; 20(3):106–111. https://pubmed.ncbi.nlm.nih.gov/9061863/, doi: 10.1016/S0166-2236(96)01002-8.

Grossberg S, Zajac L. How Humans Consciously See Paintings and Paintings Illuminate How Humans See. Art and Perception. 2017; 5(1):1–95. doi: 10.1163/22134913-00002059.

Hayden A, Bhatt RS, Quinn PC. Perceptual organization based on illusory regions in infancy. Psychonomic Bulletin and Review. 2008 4; 15(2):443–447. /record/2008-05631-030, doi: 10.3758/PBR.15.2.443.

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer Vision. 2015; 2015 Inter:1026–1034. doi: 10.1109/ICCV.2015.123.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem IEEE Computer Society; 2016. p. 770–778. doi: 10.1109/CVPR.2016.90.

Herzog MH, Ernst UA, Etzold A, Eurich CW. Local interactions in neural networks explain global effects in Gestalt masking and masking. Neural computation. 2003 9; 15(9):2091–2113. https://pubmed.ncbi.nlm.nih.gov/12959667/, doi: 10.1162/089976603322297304.

Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 .

Jacob G, Pramod RT, Katti H, Arun SP. Qualitative similarities and differences in visual object representations between brains and deep networks. Nature Communications 2021 12:1. 2021 3; 12(1):1–14. https://www.nature.com/articles/s41467-021-22078-3, doi: 10.1038/s41467-021-22078-3.
Kauderer-Abrams E. Quantifying Translation-Invariance in Convolutional Neural Networks. arXiv preprint arXiv: 180101450v1. 2017 12; http://arxiv.org/abs/1801.01450.

Kayhan OS, van Gemert JC. On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) No. class 1; 2020. p. 14262–14273. doi: 10.1109/cvpr42600.2020.01428.

Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014 11; 10(11):e1003915. doi: 10.1371/journal.pcbi.1003915.

Kim B, Reif E, Wattenberg M, Bengio S, Mozer MC. Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure. Computational Brain and Behavior. 2021 9; 4(3):251–263. https://link.springer.com/article/10.1007/s42113-021-00100-7, doi: 10.1007/S42113-021-00100-7/FIGURES/8.

Kriegeskorte N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annual Review of Vision Science. 2015 11; 1(1):417–446. www.annualreviews.org, doi: 10.1146/annurev-vision-082114-035447.

Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems. 2012; 25.

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, Majaj NJ, Issa EB, Bashivan P, Prescott-Roy J, Schmidt K, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). 2019; .

Linsley D, Kim J, Veerabradan V, Windolf C, Serre T. Learning long-range spatial dependencies with horizontal gated recurrent units. Advances in Neural Information Processing Systems. 2018; 2018-Decem(NeurIPS):152–164. doi: 10.32470/ccn.2018.1116-0.

Lotter W, Kreiman G, Cox D. A neural network trained for prediction mimics diverse features of biological neurons and perception. Nature Machine Intelligence. 2020; 2(4):210–219. http://dx.doi.org/10.1038/s42256-020-0170-9, doi: 10.1038/s42256-020-0170-9.

Malhotra G, Dujmović M, Bowers JS. Feature blindness: a challenge for understanding and modelling visual object recognition. PLOS Computational Biology. 2022; doi: 10.1101/2021.10.20.465074.

Malhotra G, Evans BD, Bowers JS. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. Vision Research. 2020 9; 174:57–68. doi: 10.1016/J.VISRES.2020.04.013.

Metzger W. Handbuch der Psychologie 1. Band 1. Halbband“ – Bücher gebraucht, antiquarisch & neu kaufen. Gottingen: Hogrefe; 1966.

Pang Z, O'may C, Choksi B, Vanrullen R. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. arXiv preprint arXiv: 210201955v2. 2021; .

Perrett DI, Oram MW. Neurophysiology of shape processing. Image and Vision Computing. 1993; 11(6):317–333. doi: 10.1016/0262-8856(93)90011-5.

Peterson MA. Past experience and meaning affect object detection: A hierarchical Bayesian approach. Psychology of Learning and Motivation - Advances in Research and Theory. 2019 1; 70:223–257. doi: 10.1016/BS.PLM.2019.03.006.

Peterson MA, Gibson BS. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. Perception & Psychophysics 1994 56:5. 1994 9; 56(5):551–564. https://link.springer.com/article/10.3758/BF03206951, doi: 10.3758/BF03206951.

Pomerantz JR, Pristach EA. Emergent features, attention, and perceptual glue in visual form perception. Journal of Experimantal Psychology: Human Perception and Perormance. 1989; 15(4):635–649.

Pomerantz JR, Portillo MC. Grouping and Emergent Features in Vision: Toward a Theory of Basic Gestalts. Journal of Experimental Psychology: Human Perception and Performance. 2011 10; 37(5):1331–1349. /record/2011-13455-001, doi: 10.1037/A0024330.

Pomerantz JR, Sager LC, Stoever RJ. Perception of wholes and of their component parts: Some configural superiority effects. Journal of Experimental Psychology: Human Perception and Performance. 1977; 3(3):422–435. doi: 10.1037/0096-1523.3.3.422.
Experiment 1. Additional Results

In Experiment 1 in the main text we showed that networks produced high CSE for 3 stimuli out of 4 only at the last fully connected layer, and no CSE at the last convolutional layer. Here we show that the same lack of CSE is observed across all earlier convolutional layers. For each network, we took the convolutional layer at 1/4 of the whole networks depth (Early Stage), 2/4 (Middle-Early) and 3/4 (Middle-Late). We plot these analysis in Figure 7, showing that in all these stages of processing networks appeared to capture crowding/interference relations (bottom-left quadrant of each subplot) but Gestalt grouping (corresponding to samples on the top-right quadrant) is almost always absent.

Appendix 0 Figure 7. Networks’ CE vs Humans’ CEs at several stages of computations, for all sets. In almost all cases, adding the contextual element does not result in facilitation (CSE) but only in interference (CIE). In fact, the inferiority effect increases at deeper layers.

Experiment 2. Additional Results

As a sanity check, we verified that the dot stimulus is sufficiently discriminable. To do that, we compare the cosine similarity between a pair of two empty canvas (that is, with random pixellated background) and a pair composed of one empty canvas and one canvas with a single dot. We observed that each networks managed to filter out the difference in random pixel across the empty canvas, as their cosine similarity appeared to be near 1 in all cases (we also verified that this is a characteristic of a trained network - vanilla cosine similarity across a pair of empty canvas was closer to 0). On the other hand, the cosine similarity between empty canvas and a single dot is much lower. This clearly indicates that a single dot is sufficiently discriminable and can be used for generating new stimuli in Experiment 2. Furthermore, the lack of sensitivity to EFs found in Experiment 2 has been tested along all networks’ depth. This is shown in Figure Figure 9, together with the plot of the difference in similarity across the pair of empty canvas and the pair with a single dot (blue line). The analysis shows that while the dots remains discriminable across most stages of processing, the additional dots in the proximity, orientation and linearity condition always hinder discrimination instead of facilitating it.
Appendix 0 Figure 8. Cosine similarity for pairs of empty canvas and pairs made of an empty canvas vs a single dot. The similarity in the second case is much lower, indicating that a single dot on a random-pixelated canvas is indeed highly discriminable. We obtained similar results with other stroke-over-background conditions.

Appendix 0 Figure 9. Amount of Configural Effect across networks’ layers. The “Single Dot” condition refers to the CSE produced by adding a single dot to the canvas (at the last stage, this is the difference between each pair of bars in Figure 8). For most layers, the single dot condition is highly discriminable. However, an additional dot aimed at eliciting either proximity, orientation, or linearity, never produces a strong CSE. In fact, it seems to hinder discrimination across most layers of computation.