Do DNNs trained on Natural Images acquire Gestalt Properties?

Valerio Biscione[a], Jeffrey S. Bowers[a]

[a]*Department of Psychology, University of Bristol, Bristol, BS8 1TL, United Kingdom*

Abstract

Under some circumstances, humans tend to perceive individual elements as a group or 'whole'. This has been widely investigated for more than a century by the school of Gestalt Psychology, which formulated several laws of perceptual grouping. Recently, Deep Neural Networks (DNNs) trained on natural images have been proposed as compelling models of human vision based on reports that they learn internal representations similar to the primate ventral visual stream and show similar patterns of errors in object classification tasks. That is, DNNs often perform well on brain and behavioral benchmarks. Here we compared human and DNN responses in discrimination judgments that assess a range of Gestalt organization principles (Pomerantz et al., 1977; Pomerantz and Portillo, 2011). Amongst the DNNs tested we selected models that perform well on the Brain-Score benchmark (Schrimpf et al., 2018). We found that network trained on natural images exhibited sensitivity to shapes at the last stage of classification, which in some cases matched humans responses. When shape familiarity was controlled for (by using dot patterns that would not resemble shapes) we found the networks were insensitive to the standard Gestalt principles of proximity, orientation, and linearity, which have been shown to have a strong and robust effect on humans. This shows that models that perform well on behavioral and brain benchmarks nevertheless miss fundamental principles of human vision.

Keywords: Deep Neural Networks, Gestalt grouping, Visual perception, Emergent Features

1. Introduction

Human tends to group perceptual features together in order to form a coherent whole. Understanding when this happens has been the focus of

research for Gestalt psychologists 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, more recently researchers have developed experimental designs with carefully constructed stimuli (e.g. Gabor Display, dot lattices) that allow for parametric control, richer visual displays, and objective measures of the 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 (more details in Section 1.2).

Recently there has also 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). 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 principle 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.

1.1. 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 neural representation at different levels of non-human primate ventral visual processing stream (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 score any neural networks on their similarity with the human object recognition system, 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 and Bowers, 2019), they appear to discriminate categories based on local instead of global features (Baker et al., 2018b; Malhotra et al., 2021), are much more susceptible to a low amount of image degradation (Geirhos et al., 2018), 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 on high level visual areas when DNNs' internal representation were compared to fMRI data 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 in 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., 2021, 2020). 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.

1.1.1. Neural Networks and Gestalt

Here we focus on an important set of psychological phenomena that has been explored with mixed success with DNNS, namely, Gestalt laws of organization. As far as we know, 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) disputed this conclusion, and found that several architectures (as simple as 3 layers networks trained on 3 ImageNet classes and as complex as 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 a video sequence, 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.

1.2. 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. As an example, consider a line segment as a stimulus, with its length, position, and orientation as the only distinguishable features. Adding a second line segment adds its own features to the perceptual visual field, but in addition, new features, which might be EFs depending on the relationship between the two stimuli: the angle between them, the type of intersection they form, the emergent parallelism, collinearity, symmetry, etc. Perception of 'wholes' are obtained when the new feature are salient, i.e. make it easier to identify the stimulus, as measured in a discrimination task.

In Figure 1, left, we illustrate this approach. 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* images, creating a *com*. The stimulus C is not informative by itself, but the combination of A and C might generate EFs, which can be compared with the features of the image containing B and C, which might also contain novel EFs, depending on the characteristics of the stimuli B and C. Any effect in discrimination performance between this composite pair and the base pair is therefore a measure of the effect of the EFs.

When the context changes the performance we talk about a *Configural Effect* (CE). In particular, when it facilitate performance we have a *Configural Superiority Effects* (CSE); When it hinders it, we have a *Configural Inferiority Effect* (CIE). Notice that for most contextual stimuli C, the discrimination will not be facilitated but impaired, due to several factors that include additional computational and attentional load, increase similarity, crowding. If, in spite of these effects, adding the stimulus C makes the discrimination faster, then a novel and salient EFs has emerged, and we can claim that Gestalt grouping has happened. This paradigm has been applied to an "oddity reaction time task" to directly measure the effect of several 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 approach has proven to be extremely useful in identifying what configurations give rises to EFs. Through their method, Pomerantz and colleagues have quantitatively measured CSE/CIE effects rather than rely on subjective measures. As predicted, many configurations result in CIE or very modest CSE (combination of characters, line segments forming letters, surfaces, 3D volumes); other configuration shows strong CSE (Pomerantz et al., 1977; Pomerantz and Pristach, 1989; Pomerantz and Portillo, 2011). Many of their complex stimuli are tested in Experiment 1, and in Experiment 2 we explore specific and low-level EFs with simple dot configuration following Pomerantz and Portillo (2011).

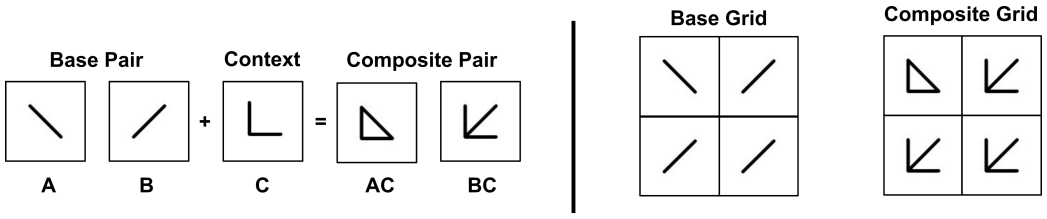

Figure 1: Pomerantz et al. (1977) approach to measure Emergent Features (EFs). **Left**: starting with a base pair, a non-informative context is added to obtain a composite pair. **Right**: The base pair and composite pair are arranged in a 2x2 grid to form an odd-discrimination task. Participants are 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.

1.3. 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 configuration 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).

On one hand, some Gestalt properties observed in 2D images have a close correspondence to analogous features of real 3D objects: elements grouped

through "closure" could be seen as corresponding to real objects partially occluded; different elements that are near and similar to each other do often correspond to variations of a texture coming from the same object (proximity, similarity), which will often move together (common fate); and so on (Todorović, 2011). This suggests that some basic Gestalt principle could be the result of applying statistical regularities acquired in everyday life, and recently some evidence has emerged in support of this hypothesis: Peterson and Gibson (1994a) 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 training (a clear example of that is given by the Kanizsa stratification images).

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. By failing to find grouping principles in our experiments we could infer that training on a more complex dataset (e.g. an interactive 3D world) might be necessary for some Gestalt principles to emerge.

2. Outline of the current work

We test a wide variety of DNNs on several sets of stimuli. Each set is composed by 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 difference 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 CSE/CIE found in humans participants, assessed through reaction times (RTs) recording (more details in Section 3.3). 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. (see Section 1.2).

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 (see Section 3.1). The human RTs data were collected from two sources: Pomerantz et al. (1977) and Pomerantz and Portillo (2011).

For Experiment 1, we used a highly varied set of stimuli from Pomerantz 1977. We compared networks' CEs to humans' CEs on 17 sets 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 Section 5.1. The structure of each set allowed for the investigation of specific emergent features, excluding confounding effects such as shape familiarity. 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.

3. Methods

3.1. Network Used

We selected 8 networks based on their historical importance, their performance on standard dataset, and their biological plausibility. We use as a point of reference the Brain-Score value (Schrimpf et al., 2018), indicating the amount of variance explained by the model across several benchmark. **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 particularly 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 (Densenet-121) has been showed to possess human like translation invariance (Biscione and Bowers, 2021).

We also tested two model 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 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 ~ 0.45 (the highest scoring network obtained 0.468[1]). InceptionNet V3, Cornet-S and its VOneNet scored slightly lower (~ 0.42). Finally, amongst the model we used, AlexNet scored the lowest (0.406). All networks were pretrained on ImageNet.

3.2. 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 Section 6.1: the participants were asked to press a key on a keyboard indicating which quadrant contained the odd stim-

[1]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

ulus, 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 below). 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 particular 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:

$$C^l(\mathbf{a}, \mathbf{b}) = \frac{d^l(\mathbf{a}) \cdot d^l(\mathbf{b})}{\|d^l(\mathbf{a})\| \, \|d^l(\mathbf{b})\|},$$

where $d^l(\mathbf{a})$ is the activation at layer l given input vector $\mathbf{a}$.

We then used the cosine similarity value to measure discrimination effect due to grouping as described below.

3.3. 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:

$$NetworkCE = C^l(base_a, base_b) - C^l(composite_a, composite_b).$$

Where each arguments to C^l is an image for one of the pair (e.g. the left image of base pair in Figure 1 would be $base_a$, the right image $base_b$). 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 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 releated to interference effects (crowding, attention load), so our focus will be on CSEs.

4. Experiment 1

4.1. Stimuli

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°). The same transformation was applied to both images of each pair. For the conditions employing transformation, each pair was tested 100 times. As we will detail below, the results were highly consistent across all transformation conditions and all stroke-over-background conditions, and therefore we will only present the results for the condition with translation and the black item over a random pixel background.

4.2. Results

CSEs are a proxy for Gestalt grouping, and so we focused on the 5 sets that exhibit 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 sped up from 40% to 180%). To match human perception, all these sets should elicit large CSEs in the networks. When analysing the last fully connected layer, set 1, 2 and 3 indeed produced large CSEs for most network (the clear exception being AlexNet), with the strongest effect shown by InceptionNet for set 1 and 2. However, set 4 and 5 produced either no effect or slightly Inferiority effect (Figure 2, top). We also computed the same analysis for and earlier stage, the last convolutional laye,r finding that the CSEs for set 1 and set 2 disappeared for most networks (with the exception of InceptioNet), becoming CIEs instead (Figure 2, bottom).

We extended the analysis at the full range of sets by plotting humans CEs vs the networks CEs in 3 (right) for the last layer (a fully connected

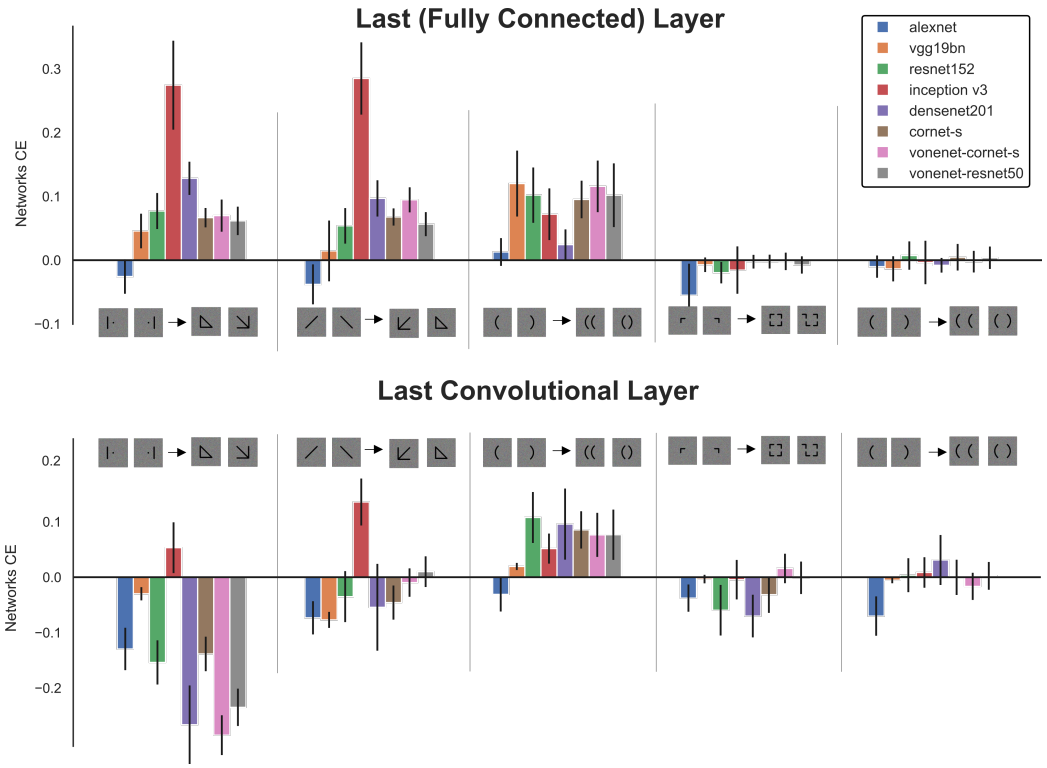

Figure 2: **Top**: Networks Configural Superiority Effects (CSEs) for the last layer (which is a fully connected layer), for the 5 sets producing high CSEs in humans. **Bottom**: At the last convolutional layer only InceptionNet exhibited the same pattern of discrimination, and some sets elicited an inferiority effect instead.

layer, circles), and the last convolutional layer (triangles).

To quantify the non-linear relationship we computed the Spearman's rank

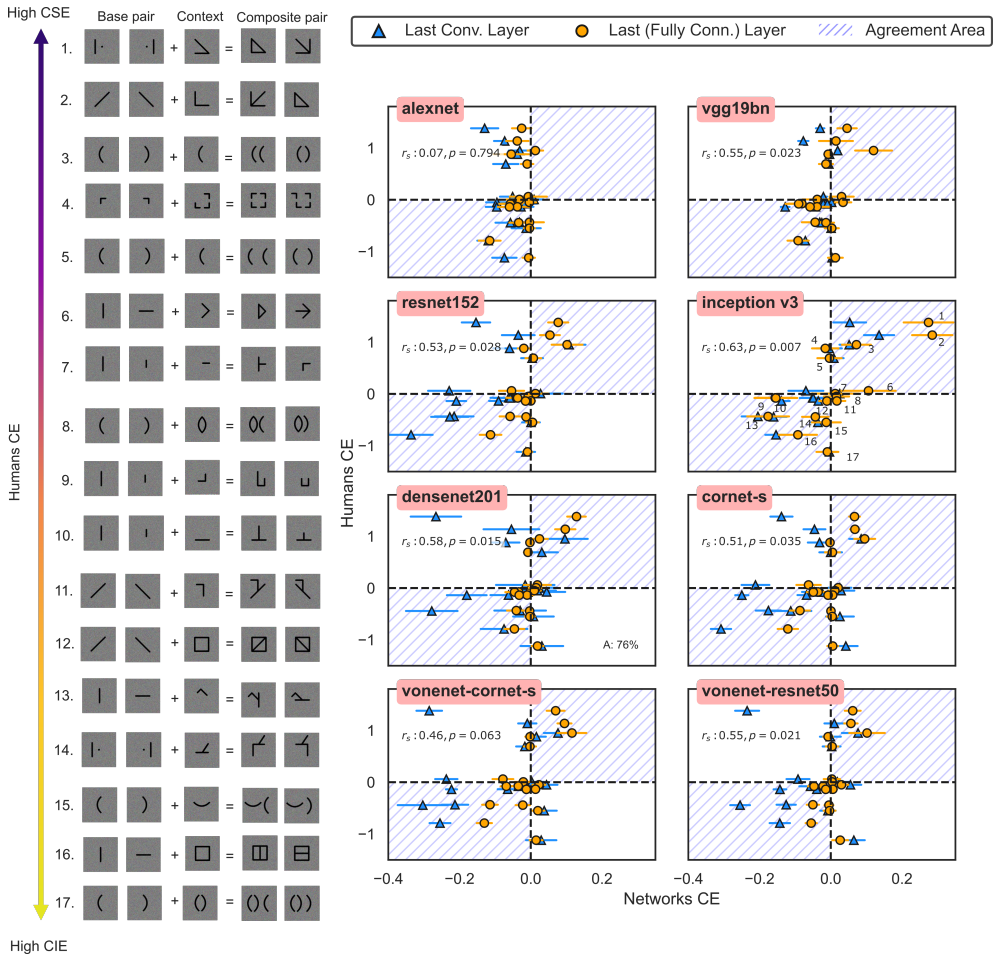

Figure 3: **Left**: the full set of base and composite pairs used in the experiment to assess Configural Effects (CE) in networks. Sets are sorted by the amount of CEs elicited in humans, measured in Pomerantz et al. (1977). Gestalt grouping is measured by CSE. **Right**: networks and humans' CE for 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)

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 $p = 0.007$). Almost the totality of the correlation is driven by 5 sets: 1, 2, and 3 (for CSE) and 13, 16 (for CIE). Repeating the analysis without these sets resulted in a much lower rank correlation (close to 0 in almost all cases) and not significant for all networks (see Appendix A).

We repeated the same analysis at earlier networks' layer. The results matched the outcome of the last convolutional layer: most of the sets producing CSEs in humans resulted in CIEs in the networks (see Appendix A).

4.3. Experiment 1. Discussion

Overall, this Experiment produced mixed results. CEs were similar across most networks excluding AlexNet (this difference is accounted for in Section 5.3). Overall, only few of the sets with high CSEs evoked similarly high CSEs in networks. Sets producing CIE in humans, indicating computational load due to distracting context, were sometimes matched by the networks, but other times resulted in weak CSEs instead. Only three out of the five CSEs sets that induced strong Gestalt grouping in humans resulted in high CSEs in the networks as well, and only at the last layer. One way to explain the pattern of agreements is to consider that the networks appeared to produce high CSEs when comparing a shape with a non-shape (set 3 could be considered a circular shape, considering that networks exhibit a certain degree of Gestalt closure Kim et al. (2021)). However, when comparing a shape with another shape (set 4) or two non-shapes (set 5) the effect disappears. This idea is consistent with the fact that CSEs were only obtained at the classification layer, but not earlier. It seems likely that by pretraining on ImageNet, the networks learnt to be sensitive to some basic shapes. Shape familiarity can indeed be an important feature for grouped perception in humans (Peterson, 2019; Duncan, 1984); however, clearly humans also use other features to group elements (Pomerantz and Portillo, 2011). With Experiment 2 we can asses whether grouping effect in networks emerges even in the absence of shape familiarity.

5. Experiment 2

Pomerantz and Portillo (2011) designed a odd-discrimination task in which dot patterns were used to create base and composite pairs of stimuli

that allowed the investigation of specific Emergent Features (EFs). In their work, CSEs of around 0.4 seconds were consistently found for three emergent features: orientation, proximity and linearity. By using their design we can not only test networks' sensitivity to these EFs, but also investigate whether networks exhibit Gestalt grouping in the absence of shape familiarity.

5.1. 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. Using the same idea illustrated in Section 1.2 we added a context canvas that would elicit either the Emergent Feature of proximity, orientation, or linearity (Figure 4). 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.

As a sanity check, we verified that the dot stimuli were salient to be discriminated by the networks. We compared the cosine similarity between a pair of empty canvas and a pair composed of an empty canvas and a canvas with a single dot. The CSEs obtained with the added dot confirmed the saliency of the stimulus. This was found across all networks, and all stages of processing, including the last fully connected layer (more details in Appendix B).

5.2. 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 RT (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 set with the base one. For a network to match human behaviour, it should produce high CSEs across all three EFs conditions.

The results are shown in Figure 5 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 at

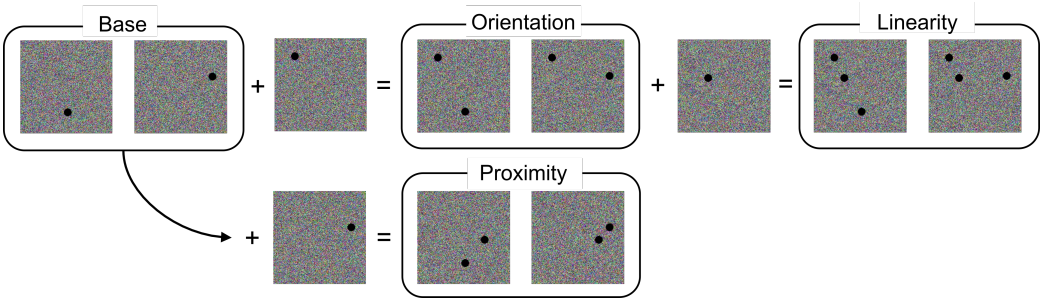

Figure 4: Generation of stimuli for Experiment 2, following Pomerantz and Portillo (2011). Starting with a pair of images in which the only discriminant feature is the location, an additional dot is added, yielding the EF of proximity or orientation. The EF of linearity is obtained by adding a dot to 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).

.

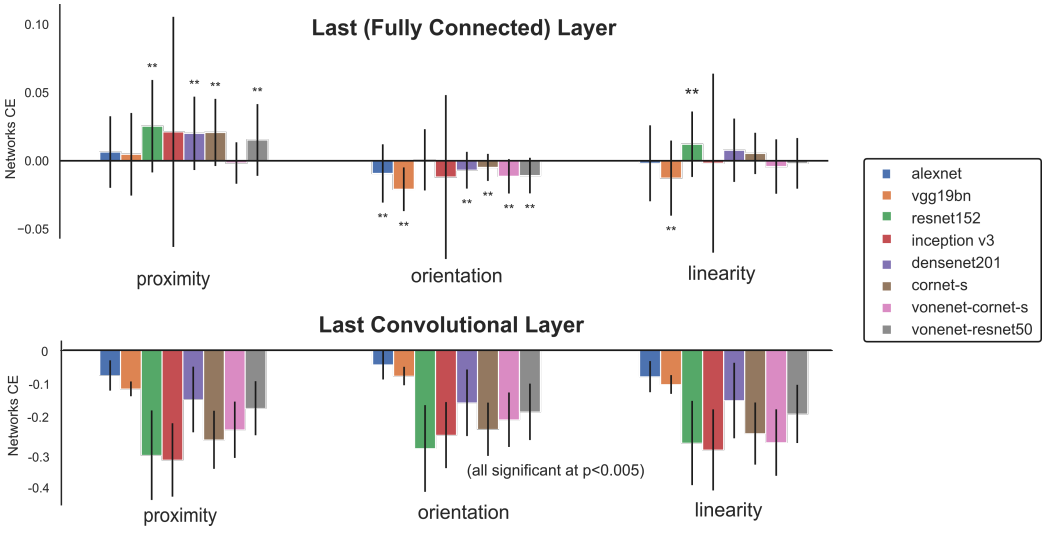

Figure 5: Amount of Configural Effect (CE) in Networks. In humans, all three of these conditions produced high CSE. In networks, the CSE was significant only for proximity, with an extremely small effect size, and only in the last fully connected layer (**top**). On earlier stage, only the additional dot was hindering discrimination instead than facilitating it (**bottom**, see also Figure B.9 in Appendix B).

the base pair. Being the cosine similarity of the base pair equal to 0.9^2 we

[2]this high value is explained by considering that the base pair contains two images

can be sure that there is plenty of space for this metric to decrease, and the minuscule effect size observed is not an artifact of the metric.

Across all other layers, the models appeared to be sensitive to additional features *in the opposite way* as humans (Figure 5, bottom, for the last convolutional layer, but similar result were observed across the whole networks' depth, see Appendix B). 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.

5.3. Are the networks capable of learning orientation, proximity and linearity?

One explanation for the lack of sensitivity to emergent features could be that the networks do not have the computational capacity to learn these features, regardless of the training dataset. In order to test this hypotheses, we directly trained each network to distinguish these features. We used the same types of dots stimuli used in the previous analysis. 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° and 25°, between 35° and 60°, and between 75° and 90°). 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. Almost all networks quickly learnt to classify based on proximity, orientation or linearity, reaching an accuracy of around 99% on the test set. Interestingly, AlexNet did not learn any of the three tasks, staying at chance level across the whole training session. Based on these results we excluded the possibility that the networks (other than AlexNet) were not architecturally able to acquire these properties.

---

which differs only on dot location. The networks tested here were pretrained on ImageNet which is shown by Biscione and Bowers 2020 to provide the networks with translation tolerance, a property also possessed by humans (Blything et al., 2021). This produced similar internal representation across different locations, and thus high cosine similarity.

6. 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.

6.1. Do modern DNNs exhibit Gestalt grouping principles?

Our findings suggest that DNNs are only very midly sensitive to proximity and not sensitive to the other Emergent Features (EFs) that several studies have studied in humans. 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). Given that Gestalt grouping principles are thought to play a key role in visual perception and object recognition, this is an important difference between how current DNNs (including DNNs that are claimed to perform well on Brain-Score) and human vision works.

In Experiment 1 the direct analysis of stimuli evoking strong Gestalt grouping produced mixed result, with only 3 of the 5 sets generating strong Configural Superiority Effect (CSEs). Furthermore, the moderate rank-correlation found across all stimuli at the last layer seemed to be completely driven by few stimuli (see Section A.6), and it disappeared at the earlier convolutional layers (see Figures 2 and Section Appendix A). Given EFs (and Gestalt rules more generally) in humans are thought to support the figure-ground segregation (Wagemans et al., 2012a) and building representations of object parts (Biederman, 1987), EFs should not be restricted to the final layer of DNNs involved in identifying images. Indeed, there is evidence that Gestalt processes occur in early visual areas (Alexander and Van Leeuwen, 2010).

Interestingly, the subset stimuli that drove the CSEs in the final layer of some DNNs could be explained by network sensitivity to shapes such as triangles, circles, etc., suggesting that it was the encoding of familiar visual features rather than EFs that led to greater discriminability of those patterns. Consistent with this, we found no evidence for strong EFs for dot stimuli in Experiment 2 at any layer of the network. The dot stimuli are less similar to familiar stimuli, and nevertheless, they generated strong EFs in human participants. It is possible that shape familiarity plays a role in grouping for humans as well (Peterson and Gibson, 1994b; Duncan, 1984), and it is

interesting that this property can emerge by training on 2D natural images that do not contain those shapes, but apparently this is not sufficient to acquire sensitivity to lower-level features.

6.2. 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. 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 (stimuli that looked most like familiar objects), and only for the last fully connected layer (with the exception of InceptionNet). Most networks showed very mild or non-existent effect on proximity, and null or negative effect of orientation and linearity in Experiment 2. An interesting outlier was AlexNet, which showed much smaller CSE/CIE in Experiment 1 and indeed failed to explicitly learn relevant features in Section 5.3. The similar outcomes across most networks suggests that Brain-Score rankings are not capturing a fundamental feature of human vision. This adds to a long list of other important psychological findings that the DNNs with the highest brain-scores fail to explain (Bowers et al., 2022).

6.3. 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 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 shapes (Experiment 1), regardless of the background their are presented to (white, black, or random pixels). This is impressive considering that ImageNet contained naturalistic images that are very dissimilar to the abstract shapes the network showed sensitivity to. But the critical point for our purposes is that training on ImageNet did only induce very weak EFs for proximity, and none for orientation and linearity that are so salient for humans.

However, our experiment does not show that grouping property cannot be extracted from statistical regularities of the visual environment; instead,

it shows that even such a large dataset of 2D shapes is not enough to acquire these properties. It 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, arguing that in DNNs the basic grouping properties tested here are not emerging from learning on a naturalistic 2D dataset does not imply that other properties such as closure could not be. Gestalt properties are not acquired all together in one go, 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)). This does not incontrovertibly indicates a gradual acquisition based on learning, but it might depend on the on-going development of the visual system in infants. In fact, the work Kim et al. (2021) seem to indicate that networks trained on ImageNet exhibit closure (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.

7. Conclusion

Overall, the results presented in this work cast some doubt on the current possibility of using DNNs trained on natural 2D images as a model of the human visual system for object recognition. In spite of the success on benchmarks such as Brain-Score, the experiments indicated a general disagreements in discrimination judgments between networks and humans on pair of stimuli aimed to elicit Gestalt grouping effects. Noteworthy, similarly to human participants, networks seemed sensitive to stimuli containing familiar shapes, even though these were not present in the training set (Experiment 1). When this was controlled for in a task aimed to measure the

emergent features of proximity, orientation, and linearity (Experiment 2), networks did not show any human-like Gestalt effect, and in fact these features *hindered* discrimination. We further tested whether this was due to an inability to learn such properties, and we found it not to be the case for any network but AlexNet, which did not seem computationally able to capture the tested Gestalt principles. 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.

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).

Appendix A. Experiment 1. Additional Results

Most of the rank-correlatoin found in Experiment 1 across human and networks CEs is fully driven by 5 sets: 1, 2, 3, 13, and 16. Here we show the ρ values when this sets are excluded (Figure A.6).

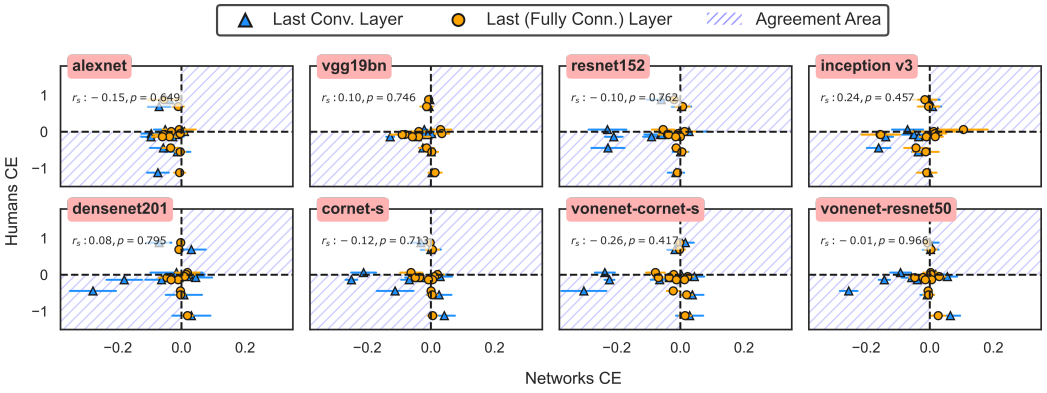

Figure A.6: **A**: blalba

As shown in the Figure, the correlation becomes low and non-significant for all networks.

We also computed the same analysis across many layers of each network in the following way: for each network, we took the convolutional layer at one forth of the whole networks depth (Early Stage), 2/4 (Middle-Early) and 3/4 (Middle-Late). In all cases, these corresponded to convolutional layers (the analysis of later layers are shown in the main text). We plot these analysis in Figure A.7, showing that in all these stages of processing. As usual, network appear to capture crowding/interference relations (bottom-left quadrant of each subplot) but across all stages Gestalt grouping (corresponding to samples on the top-right quadrant) is almost always absent.

Appendix B. 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

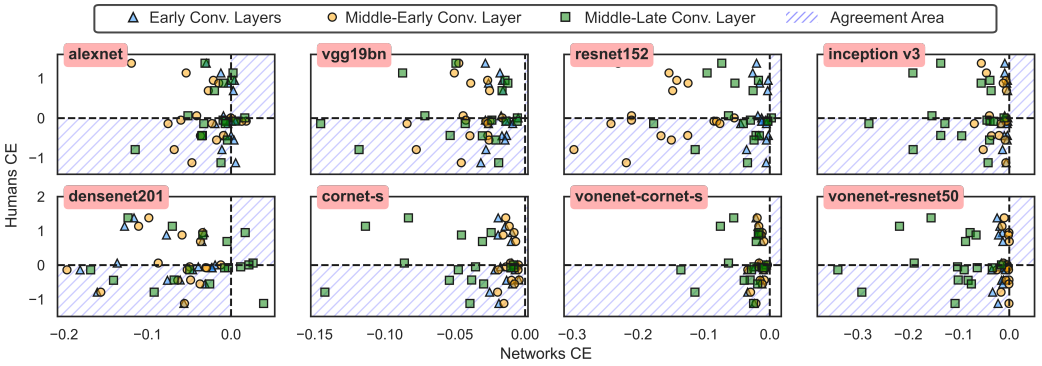

Figure A.7: **A**: blalba

(we also verified that this is a characteristic of a trained network - vanilla cosine similarity across a pair of empty canvas was closer to 0.2). 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,

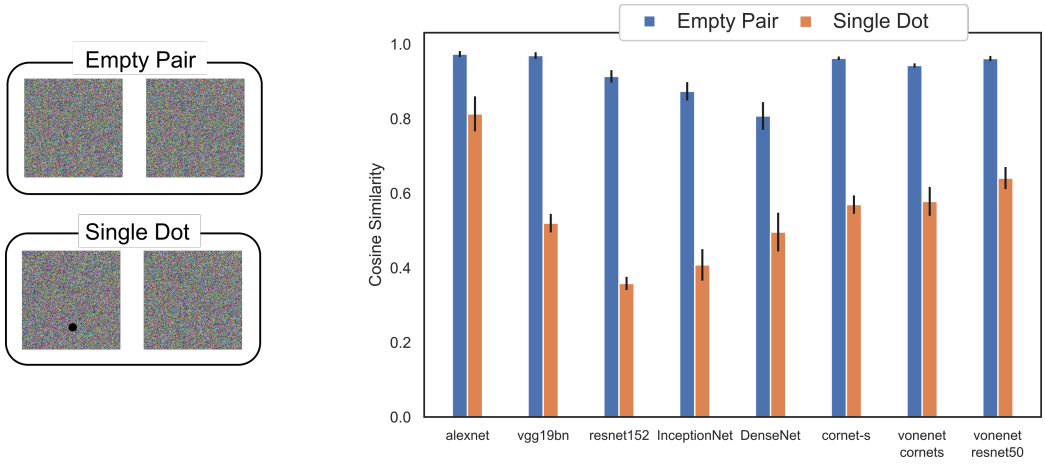

Figure B.8: **A**: blalba

the lack of sensitivity to EFs found in Experiment 2 has been tested along all networks' depth. This is shown in Figure B.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.

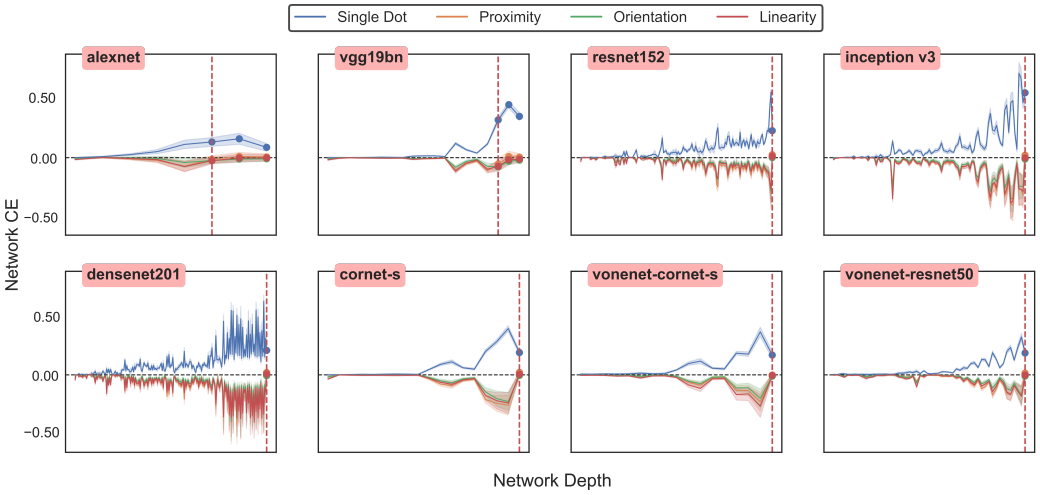

Figure B.9: **A**: blalba

References

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

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

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

Biederman, I., 1987. Recognition-by-Components: A Theory of Human Image Understanding. Technical Report 2.

Biscione, V., Bowers, J., 2020. Learning Translation Invariance in CNNs. 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2020 URL: http://arxiv.org/abs/2011.11757.

Biscione, V., Bowers, J.S., 2021. Convolutional Neural Networks Are Not Invariant to Translation, but They Can Learn to Be. Journal of Machine Learning Research 22, 1–28. URL: http://jmlr.org/papers/v22/21-0019.html.

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

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

Blything, R., Biscione, V., Vankov, I.I., Ludwig, C.J.H., Bowers, J.S., 2021. The human visual system and CNNs can both support robust online translation tolerance following extreme displacements. Journal of Vision 21, 1–16. URL: https://doi.org/10.1167/jov.21.2.9., doi:10.1167/jov.21.2.9.

Chen, F.X., Roig, G., Isik, L., Boix, X., Poggio, T., 2017. Eccentricity dependent deep neural networks: Modeling invariance in human vision. AAAI Spring Symposium - Technical Report SS-17-01 -, 541–546.

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

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

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

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

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

Geirhos, R., Medina Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A., Temme, C.R.M., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A., 2018. Generalisation in humans and deep neural networks, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 7538–7550.

German, J.S., Jacobs, R.A., 2020. Can machine learning account for human visual object shape similarity judgments? Vision Research 167, 87–99. URL: https://doi.org/10.1016/j.visres.2019.12.001, doi:10.1016/j.visres.2019.12.001.

Gong, Y., Wang, L., Guo, R., Lazebnik, S., 2014. 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) 8695 LNCS, 392–407. URL: http://arxiv.org/abs/1403.1840.

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

Hayden, A., Bhatt, R.S., Quinn, P.C., 2008. Perceptual organization based on illusory regions in infancy. Psychonomic Bulletin and Review 15, 443–447. URL: /record/2008-05631-030, doi:10.3758/PBR.15.2.443.

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

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

Herzog, M.H., Ernst, U.A., Etzold, A., Eurich, C.W., 2003. Local interactions in neural networks explain global effects in Gestalt processing and masking. Neural computation 15, 2091–2113. URL: https://pubmed.ncbi.nlm.nih.gov/12959667/, doi:10.1162/089976603322297304.

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

Jacob, G., Pramod, R.T., Katti, H., Arun, S.P., 2021. Qualitative similarities and differences in visual object representations between brains and deep networks. Nature Communications 2021 12:1 12, 1–14. URL: https://www.nature.com/articles/s41467-021-22078-3, doi:10.1038/s41467-021-22078-3.

Kauderer-Abrams, E., 2017. Quantifying Translation-Invariance in Convolutional Neural Networks. arXiv preprint arXiv: 1801.01450v1 URL: http://arxiv.org/abs/1801.01450.

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

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

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

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

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

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

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

Malhotra, G., Bowers, J., 2019. The contrasting roles of shape in human vision and convolutional neural networks. Proceedings of the 41st Annual Conference of the Cognitive Science Society 2019 URL: https://pillow.readthedocs.io.

Malhotra, G., Dujmović, M., Bowers, J.S., 2021. Feature blindness: a challenge for understanding and modelling visual object recognition. bioRxiv , 2021.10.20.465074URL: https://www.biorxiv.org/content/10.1101/2021.10.20.465074v2https://www.biorxiv.org/content/10.1101/2021.10.20.465074v2.abstract, doi:10.1101/2021.10.20.465074.

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

Pang, Z., Biggs O'may, C., Choksi, B., Vanrullen, R., 2021. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. arXiv preprint arXiv: 2102.01955v2 .

Perrett, D.I., Oram, M.W., 1993. Neurophysiology of shape processing. Image and Vision Computing 11, 317–333. doi:10.1016/0262-8856(93)90011-5.

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

Peterson, M.A., Gibson, B.S., 1994a. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. Perception & Psychophysics 1994 56:5 56, 551–564. URL: https://link.springer.com/article/10.3758/BF03206951, doi:10.3758/BF03206951.

Peterson, M.A., Gibson, B.S., 1994b. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. Perception & Psychophysics 1994 56:5 56, 551–564. URL: https://link.springer.com/article/10.3758/BF03206951, doi:10.3758/BF03206951.

Pomerantz, J.R., Portillo, M.C., 2011. Grouping and Emergent Features in Vision: Toward a Theory of Basic Gestalts. Journal of Experimental Psychology: Human Perception and Performance 37, 1331–1349. URL: /record/2011-13455-001, doi:10.1037/A0024330.

Pomerantz, J.R., Pristach, E.A., 1989. Emergent features, attention, and perceptual glue in visual form perception. Journal of Exerpimental Psychology: Human Perception and Perormance 15, 635–649.

Pomerantz, J.R., Sager, L.C., Stoever, R.J., 1977. Perception of wholes and of their component parts: Some configural superiority effects. Journal of Experimental Psychology: Human Perception and Performance 3, 422–435. doi:10.1037/0096-1523.3.3.422.

Puebla, G., Bowers, J.S., 2021. Can deep convolutional neural networks support relational reasoning in the same-different task? bioRxiv , 2021.09.03.458919URL: https://www.biorxiv.org/content/10.1101/2021.09.03.458919v1https://www.biorxiv.org/content/10.1101/2021.09.03.458919v1.abstract, doi:10.1101/2021.09.03.458919.

Quinn, P.C., Bhatt, R.S., 2005. Learning perceptual organization in infancy. Psychological Science 16, 511–515. doi:10.1111/j.0956-7976.2005.01567.x.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D.L.K., DiCarlo, J.J., 2018. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv , 407007URL: https://www.biorxiv.org/content/10.1101/407007v1, doi:10.1101/407007.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556 URL: http://www.robots.ox.ac.uk/.

Spillmann, L., 2009. Phenomenology and neurophysiological correlations: Two approaches to perception research. Vision Research 49, 1507–1521. doi:10.1016/J.VISRES.2009.02.022.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June, 1–9. URL: https://arxiv.org/abs/1409.4842v1, doi:10.1109/CVPR.2015.7298594.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings URL: https://arxiv.org/abs/1312.6199v4.

Todorović, D., 2011. What is the Origin of the Gestalt Principles. Humanamente 17, 1–20.

Treisman, A., 1998. Feature binding, attention and object perception. Philosophical Transactions of the Royal Society B: Biological Sciences 353, 1295. URL: /pmc/articles/PMC1692340/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1692340/, doi:10.1098/RSTB.1998.0284.

Vecera, S.P., Farah, M.J., 1997. Is visual image segmentation a bottom-up or an interactive process? Perception & psychophysics 59, 1280–1296. URL: https://pubmed.ncbi.nlm.nih.gov/9401461/, doi:10.3758/BF03214214.

Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M., von der Heydt, R., 2012a. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychological Bulletin 138, 1172–1217. doi:10.1037/a0029333.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J.R., Van der Helm, P.A., Van Leeuwen, C., 2012b. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. Psychological Bulletin 138, 1218–1252. doi:10.1037/a0029334.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., Tanaka, K., 2018. Illusory motion reproduced by deep neural networks trained for prediction. Frontiers in Psychology 9, 345. doi:10.3389/FPSYG.2018.00345/BIBTEX.

Wertheimer, M., 1923. Untersuchungen zur Lehre von der Gestalt. II. Psychologische Forschung 1923 4:1 4, 301–350. URL: https://link.springer.com/article/10.1007/BF00410640, doi:10.1007/BF00410640.

Xu, Y., Vaziri-Pashkam, M., 2021. Examining the Coding Strength of Object Identity and Nonidentity Features in Human Occipito-Temporal Cortex and Convolutional Neural Networks. The Journal of Neuroscience

41, 4234–4252. URL: https://pubmed.ncbi.nlm.nih.gov/33789916/, doi:10.1523/jneurosci.1993-20.2021.

Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience 2016 19:3 19, 356–365. URL: https://www.nature.com/articles/nn.4244, doi:10.1038/nn.4244.

Zemel, R.S., Mozer, M.C., Behrmann, M., Bavelier, D., 2002. Experience-dependent perceptual grouping and object-based attention. Journal of Experimental Psychology: Human Perception and Performance 28, 202–217. doi:10.1037/0096-1523.28.1.202.