
Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure

Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, Michael C. Mozer
Google Research
1600 Amphitheater Parkway
Mountain View, CA 94043

The Gestalt laws of perceptual organization, which describe how visual elements in an image are grouped and interpreted, have traditionally been thought of as innate despite their ecological validity. We use deep-learning methods to investigate whether natural scene statistics might be sufficient to derive the Gestalt laws. We examine the law of *closure*, which asserts that human visual perception tends to “close the gap” by assembling elements that can jointly be interpreted as a complete figure or object. We demonstrate that a state-of-the-art convolutional neural network, trained to classify natural images, exhibits closure on synthetic displays of edge fragments, as assessed by similarity of internal representations. This finding provides support for the hypothesis that the human perceptual system is even more elegant than the Gestaltists imagined: a single law—adaptation to the statistical structure of the environment—might suffice as fundamental.

Psychology has long aimed to discover fundamental laws of behavior that place the field on the same footing as ‘hard’ sciences like physics and chemistry (Schultz and Schultz, 2015). Perhaps the most visible and overarching set of such laws, developed in the early twentieth century to explain perceptual and attentional phenomena, are the *Gestalt principles* (Wertheimer, 1923). These principles have had a tremendous impact on modern psychology (Kimchi, 1992; Wagemans et al., 2012a,b; Schultz and Schultz, 2015). Although Gestalt psychology has faced some criticism over a lack of rigor (Wagemans et al., 2012a; Westheimer, 1999; Schultz and Schultz, 2015), investigators have successfully operationalized its concepts (Ren and Malik, 2003), and it has influenced work in medicine (Bender, 1938), computer vision (Desolneux et al., 2007), therapy (Zinker, 1977), and design (Behrens, 1998).

The Gestalt principles describe how visual elements are grouped and interpreted. For example, the Gestalt principle of *closure* asserts that human visual perception tends to “close the gap” by grouping elements that can jointly be interpreted as a complete figure or object. The principle thus provides a basis for predicting how viewers will parse and understand display fragments such as those in Figures 1a, b. The linking of fragments such as those in Figure 1a occurs early in perceptual processing, hampering access to the constituent fragments but facilitating rapid recognition of the completed shape (Rensink and Enns, 1998).

The Gestalt principles can support object perception by grouping together strongly interrelated features—features likely to belong to the same object, allowing features of that object to be processed apart from the features of other objects (e.g., Figure 1c). Consistent with this role of grouping, the Gestalt principles have long been considered to have ecological validity in the natural world (Brunswik and Kamiya, 1953). That is, natural image statistics have been shown to justify many of the Gestalt principles, including good continuation, proximity, and similarity (Elder and Goldberg, 2002; Geisler et al., 2001; Krüger, 1998; Sigman et al., 2001).

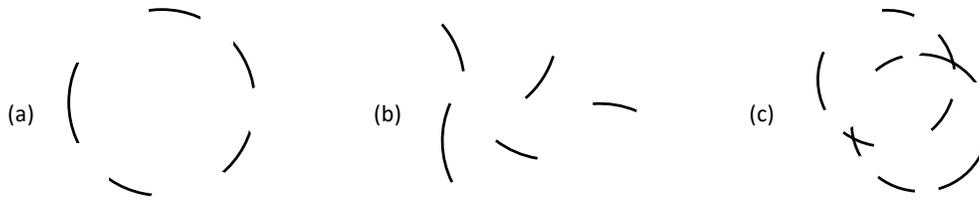


Figure 1: (a) A circle formed from fragments via closure; (b) the same fragments but rearranged to prevent closure; (c) fragments from two circles which can be segmented using closure

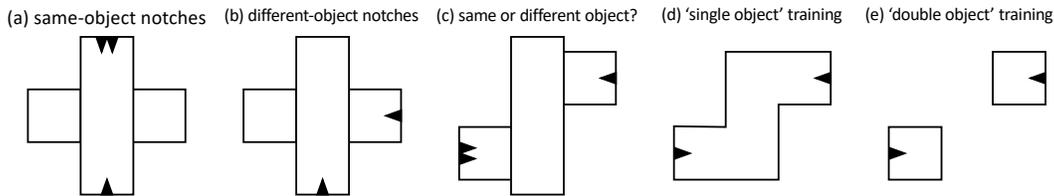


Figure 2: Examples of stimuli used by Zemel et al. (2002)

The Gestaltist tradition considered the principles to be innate and immutable (Kimchi, 1992). Although the role of learning was acknowledged, the atomic Gestalt principles were considered primary (Todorovic, 2008). Even the aforementioned research examining natural image statistics has presumed that the principles either evolved to support ordinary perception or fortuitously happen to have utility for perception.

However, ample evidence supports the notion that perceptual grouping can be modulated by experience. For example, figure-ground segregation is affected by object familiarity: a silhouette is more likely to be assigned as the figure if it suggests a common object (Peterson and Gibson, 1994; Peterson, 2019). And perceptual grouping can be altered with only a small amount of experience in a novel stimulus environment (Zemel et al., 2002). In Zemel et al.’s studies, participants were asked to report whether two features in a display matched. Consistent with Duncan (1984), participants are faster to respond when the two features—notches on the ends of rectangles—belong to the same object (Figure 2a) relative to when they belong to different objects (Figure 2b). Although participants treat Figures 2a,b as one rectangle occluding another, the two small squares in Figure 2c are treated as distinct objects. However, following brief training on stimuli such as the zig-zag shape in Figure 2d, the two small squares are treated as parts of the same object, relative to a control condition in which the training consisted of fragments as in Figure 2e.

If perceptual grouping can be modulated by experience, perhaps the Gestalt principles are not innate and immutable but rather are developmentally acquired as a consequence of interacting with the natural world. Ordinary perceptual experience might suffice to allow a learner to discover the Gestalt principles, given that the statistical structure of the environment is consistent with the Gestalt principles (Elder and Goldberg, 2002; Geisler et al., 2001; Krüger, 1998; Sigman et al., 2001). In the present work, we use deep-learning methods to investigate this hypothesis.

We focus on closure (Figure 1a,c). Closure is a particularly compelling illustration of the Gestalt perspective because fragments are assembled into a meaningful configuration and perceived as a unified whole (Wertheimer, 1923). In this process, missing fragments can be filled in, resulting in illusory contours, the classic example of which is the Kanizsa triangle (Figure 3a). Traditional cognitive models have been designed to explain illusory contours (e.g., Grossberg, 2014; Kalar et al., 2010). These models adopt the assumption of innateness in that they are built on specialized mechanisms designed to perform some form of filling in. We examine whether a deep neural net trained on natural images exhibits closure effects naturally and as a consequence of its exposure to its environment.

Closure has been examined experimentally in humans via measures that include response latency (Kimchi et al., 2016), discrimination ability (Ringach and Shapley, 1996), and EEG response (Sanguinetti et al., 2016). Measuring closure effects via behavior in feedforward neural networks is not entirely straightforward. These networks have no temporal dynamics of response formation, so

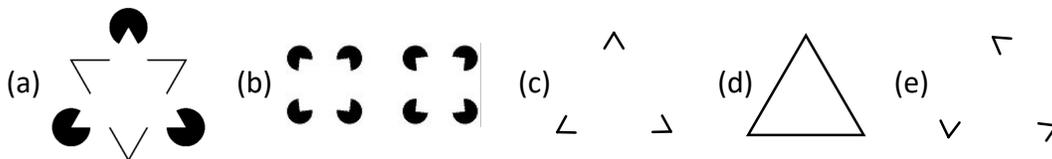


Figure 3: (a) The Kanisza triangle formed from illusory contours; (b) fat and thin squares used as stimuli by Baker et al. (2018); (c) *aligned* corners—the minimal visual cues required to induce closure and an illusory triangle; (d) a *complete* triangle; and (e) *disorderd* corners, which should be insufficient to induce closure or an illusory triangle.

latency-based response measures will be uninformative. And for discrimination tasks, they require task-specific training, which complicates the use of existing, pretrained models.

Baker et al. (2018) explored whether neural nets ‘perceive’ illusory contours using an indirect technique. They studied displays consisting of fragments that could be completed as either fat or thin rectangles (Figure 3b, left and right images, respectively). Using AlexNet (Krizhevsky et al., 2012), a pretrained network for image classification, they decapitated the output layer which represents every object class and replaced it with a single unit that discriminates fat from thin rectangles. The weights from the penultimate layer to the output unit were trained on complete (non-illusory) fat and thin rectangles presented in varying sizes, aspect ratios, and positions in the image. This additional training extracts information available from the original model for fat versus thin classification. Following training, the network could readily discriminate fat and thin rectangles, whether real or illusory. Baker et al. (2018) then borrowed a method from the human behavioral literature, *classification images* (Gold et al., 2000), to infer features in the image that drive responses. Essentially, the method adds pixelwise luminance noise to images and then uses an averaging technique to identify the pixels that reliably modulate the probability of a given response. In humans, this method infers the illusory contours of the rectangles suggested by the stimuli in Figure 3b. In contrast, Baker et al. (2018) found no evidence that that pixels along illusory contours influence network classification decisions.

Although the classification-image paradigm finds a measurable difference between networks and humans, Baker et al.’s claim that ‘deep convolutional networks do not perceive illusory contours’ (the title of their article) is too strong. Baker et al. treat networks as black boxes, which conflicts with their stated goal of probing the nature of internal representations. One advantage of network experiments over human experiments is that representations can be observed directly. With humans, the classification-image paradigm is a necessary and clever means of reverse-engineering representations; however, the paradigm recovers only *image* representations not *internal* representations. Illusory contours and filling-in processes should be manifested in internal representations. This manifestation is likely not visual in nature: Rensink and Enns (1998) found no evidence that the inferred contours are superimposed on a low-level visual representations; instead, they describe a type of functional or conceptual filling-in that influences high-level representations constructed from the low-level representation. This claim is consistent with monkey neurophysiology indicating that the first cortical stage (area 17) is not responsive to illusory contours whereas later stages (including area 18) are (von der Heydt et al., 1984).¹

In our work, we examine the internal representations of a neural network trained on natural scenes and then tested on a minimalist instantiation of the Kanisza triangle, which consists of three corner fragments with the edges between them removed (Figure 3c), similar to stimuli used in classic human studies (e.g., Elder and Zucker, 1993; Kimchi et al., 2016; Kimchi, 1992). We investigate whether these *aligned* fragments yield a representation more similar to that of a *complete* triangle (Figure 3d) than do *disordered* fragments (Figure 3e). Focusing on similarity of internal representations allows us to evaluate the Gestalt perception of shape and the functional filling in of the missing contours. The aligned fragments should induce stronger illusory contours than do the disordered fragments.

Our stimuli are pixel images of complete, aligned, and disordered shapes, varying in a number of irrelevant dimensions to ensure robustness of effects we observe (Figure 4a,b). The stimuli are

¹Regardless of the issue of where and how illusory contours are manifested in the visual stream, Baker et al. have identified a behavioral discrepancy between feedforward neural networks and humans which must be explained. We return to this issue in the discussion.

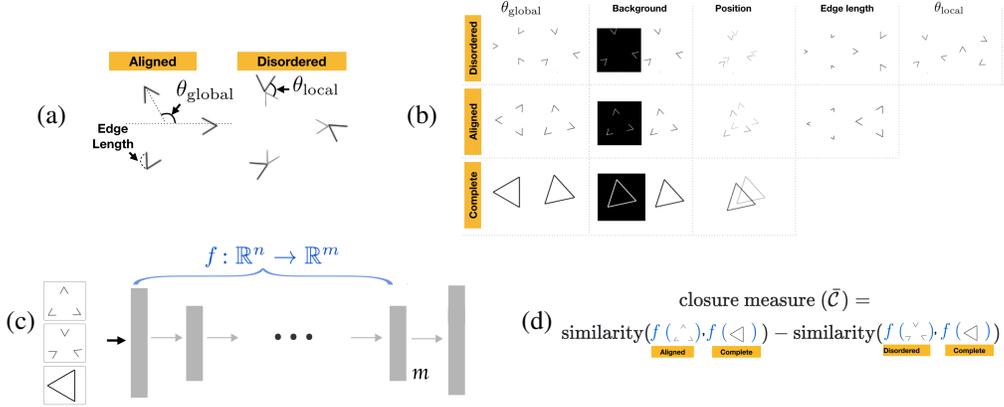


Figure 4: Outline of the stimuli and methodology to test closure in pretrained neural networks. (a) The tested shapes varied in global orientation and the disordered images also varied in local orientation of their elements. (b) Examples of stimulus variation for the three experimental conditions (depicted in the rows), and for five properties (depicted in the columns). (c) Images are fed into a deep ConvNet trained to classify images, where there is one output neuron per class. In most of our simulations, the penultimate layer, with m units is used as a deep embedding of the input image. (d) Computing a closure measure, \bar{C} , where a larger value indicates that the representation of the complete triangle is more similar to the representation of the aligned fragments than to the representation of the disordered fragments. Note that \bar{C} is an expectation over many image triples, not depicted in the equation.

fed into a pretrained deep convolutional neural net (hereafter, *ConvNet*) that maps the image to an m -dimensional internal representation, which we refer to as the *embedding* (Figure 4c). We estimate the expected relative similarity of aligned and disordered fragments to the complete image using a closure measure, $\bar{C} \in [-1, +1]$, where a larger value indicates that the representation of the complete triangle is more like the representation of the aligned fragments than the representation of the disordered fragments (Figure 4d).

Results

Sanity check

We conduct a proof-of-concept experiment to show that we can distinguish models that produce closure from those that do not. To ensure that the models have these properties, we train simple ConvNets from scratch solely on our set of complete, aligned, and disordered images. The networks are trained to perform one of two binary classification tasks: *closure discrimination (CD)*, which produces output 1 for complete and aligned images and output 0 for disordered images, and *background discrimination (BD)*, which produces output 1 for black backgrounds and 0 for white backgrounds. The CD net will necessarily treat complete and aligned as more similar than complete and disordered, and should therefore produce a positive \bar{C} score. In contrast, the BD net needs to extract color not shape information from images, and if it ignores shape altogether, it will yield a \bar{C} score of 0. Our aim in this contrast is to present the pattern of results obtained in these two conditions as signatures of closure (for CD) and lack of closure (for BD).

Figure 5 presents the closure measure (\bar{C}) for the CD and BD models, as a function of the edge length (Figure 4b). The CD model, trained to treat aligned and closure as identical, produces linearly increasing closure as the edge length increases. The BD model, trained to focus on background color, produces a flat function of closure with edge length. Thus, when a model necessarily treats aligned and complete as identical, it produces a monotonic ramp with edge length. When a model has no constraint on how it treats the different types of images, it fails to produce closure at any edge length. We therefore use these curves as signatures of closure and failure to exhibit closure, respectively.

Given that the CD model was trained to treat aligned images as equivalent to complete triangles, regardless of edge length, it is surprising that internal representations of the model remain sensitive to edge length, as evidenced by increasing closure with edge length. Because the task requires

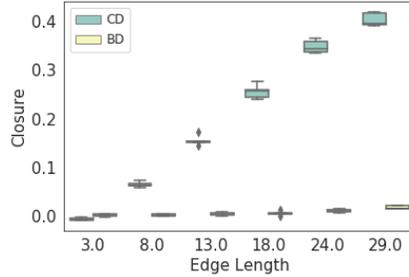


Figure 5: Sanity check experiment. CD networks, trained to discriminate complete and aligned from disordered images, show increasing closure with edge length. BD networks, trained to discriminate background color, show no closure. The shading around each line indicates a confidence interval obtained by running replications of each network. Only final layer is plotted.

determining how edges align in Gestalt shapes, it seems to be impossible for the CD model not to preserve information about edge length, albeit task irrelevant. This feature of the model is consistent with findings that human performance also varies as a continuous, monotonic function of the edge length, whether the behavioral measure of closure is discrimination threshold (Ringach and Shapeley, 1996), search latency (Elder and Zucker, 1993), or memory bias (Holmes, 1968). Similarly, neurons in area 18 of the visual cortex of alert monkeys responding to illusory contours show an increased strength of response as the edge length increases (von der Heydt et al., 1984). These empirical results give further justification to treating the profile of the CD model in Figure 5 as a signature of closure.

The role of natural image statistics

We now turn to the main focus of our modeling effort: to evaluate the hypothesis that natural image statistics, in conjunction with a convolutional net architecture, are necessary and sufficient to obtain closure in a neural net. We perform a series of simulations that provide converging evidence for this hypothesis. Each experiment compares a *base* model to a model that varies in a single aspect, either its architecture or training data. Our base model is a state-of-the-art pretrained image classification network, Inception (Szegedy et al., 2016).

Natural images versus white noise

Figure 6a compares the base model to an identical architecture trained on white noise images. The base model and white-noise net share not only the same architecture, but also training procedure and initial weight distribution; they differ only in that the white-noise net does not benefit from natural image statistics. Nonetheless, the network has the capacity to learn a training set of 1.2 million examples (same number as normal Imagenet training set) from 1001 randomly defined classes. Each pixel in these images is sampled from a uniform $[-1, +1]$ distribution, the range of values that the model has for natural images after preprocessing. The base model shows a closure effect: a monotonic increase in \bar{C} with edge length, whereas the white-noise net obtains $\bar{C} = 0$ regardless of edge length. Performing a two-way analysis of variance with stimulus triple as the between-condition random factor, we find a main effect of model ($F(1, 1188) = 2507, p < 0.0001$), a main effect of edge length ($F(5, 1188) = 254, p < 0.0001$), and an interaction ($F(5, 1188) = 254, p < 0.0001$).

Original images versus shuffled pixels

Training on white noise might be considered a weak comparison point because the low-order statistics (e.g., pairs of adjacent pixels) are quite different from those natural images. Consequently, we tested an input variant that looks quite similar to white noise to the human eye but matches pixelwise statistics of natural images: images whose pixels have been systematically shuffled between image locations. While these shuffled images contain the same information as natural images, the rearrangement of pixels not only prevents the human eye from detecting structure but also blocks the network from learning structure and regularities due to the local connectivity of the receptive fields. Nonetheless, large overparameterized neural networks like Inception have the capacity to learn the shuffled-pixel training set, although they will not generalize to new examples (Zhang et al., 2016).

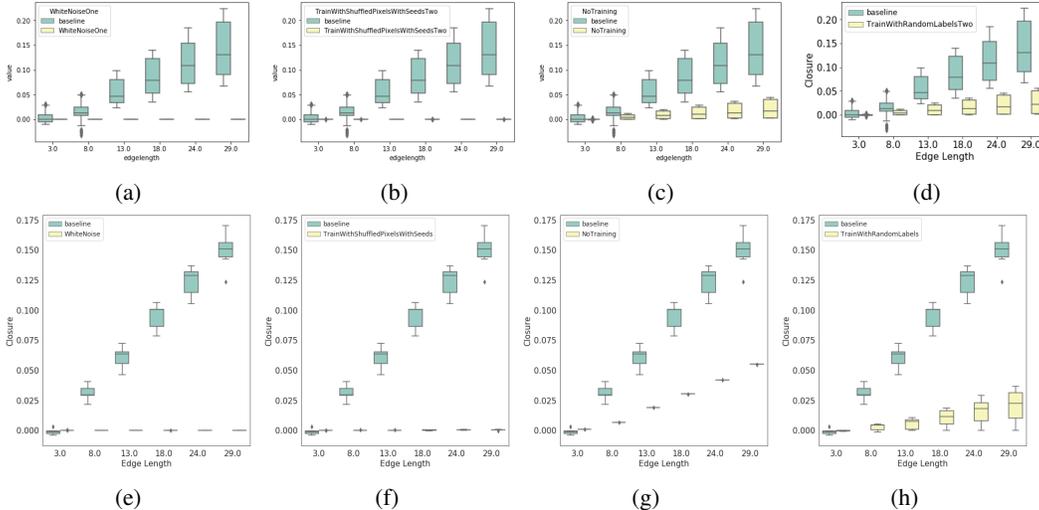


Figure 6: Exploration of how closure is influenced by various aspects of the neural net. We test Inception with 1000-classes (panels a-d) and a smaller ConvNet architecture with 3, 6, or 9 classes (panels e-h). Each graph compares a standard ConvNet architecture trained on natural images to an alternative: (a) comparing standard Inception to white-noise trained model, (b) comparing standard Inception to model trained on shuffled pixels, (c) comparing standard Inception to untrained model, (d) comparing standard Inception to model trained on shuffled labels, (e) comparing small ConvNet to white-noise trained model, (f) comparing small ConvNet to model trained on shuffled pixels, (g) comparing small ConvNet to untrained model, (h) comparing small ConvNet to model trained on shuffled labels,

Figure 6b shows that Inception trained on shuffled pixels does not obtain a closure effect. Performing a two-way analysis of variance, we find a main effect of model ($F(1, 888) = 1249.6, p < .0001$), a main effect of edge length ($F(5, 888) = 253.3, p < .0001$), and an interaction ($F(5, 888) = 126.7, p < .0001$).

Trained versus untrained networks

Our white-noise and shuffled-pixel experiments indicate that training on corrupted inputs prevents closure. Now we ask whether an untrained network naturally exhibits closure, which is then suppressed by training in the case of corrupted images.

Figure 6c compares our base model to the same model prior to training, with random initial weights. The untrained model exhibits a weaker closure effect as indicated by an interaction between condition and edge length ($F(5, 1188) = 166.9, p < .0001$). Averaging over edge lengths, the magnitude of the random-weight closure effect is nonzero ($t(599) = 19.7, p < 0.0001$), indicating that some amount of closure is attributable simply to the initial architecture and weights. This finding is not entirely surprising as researchers have noted the strong inductive bias that spatially local connectivity and convolutional operators impose on a model, making them effective as feature extractors with little or no data (Ulyanov et al., 2018; Zhang et al., 2020). In the Supplementary Materials, we show that the amount of training required for the network to reach its peak \bar{C} is fairly minimal, about 20 passes through the training set, about 1/6th of the training required for the network to reach its asymptotic classification performance.

Systematic versus shuffled labels

We have argued that the statistics of natural image data are necessary to obtain robust closure, but we have thus far not explored what aspect of these statistics are crucial. Natural image data consist of {image, label} pairs, where there is both *intrinsic* structure in the images themselves—the type of structure typically discovered by unsupervised learning algorithms—and *associative* structure in the systematic mapping between images and labels. Associative structure is crucial in order for a network to generalize to new cases.

In Figure 6d, we compare our base model with a version trained on shuffled labels, which removes the associative structure. Our model has the capacity to memorize the randomly shuffled labels, but of course it does not generalize. The shuffled-label model exhibits a weaker closure effect as indicated by an interaction between condition and edge length ($F(5, 1188) = 143.0, p < .0001$). Averaging over edge lengths, the magnitude of the shuffled-label closure effect is nonzero ($t(599) = 18.5, p < 0.0001$), indicating that some amount of closure is attributable simply to intrinsic image structure. We conjecture that the network must extract this structure in order to compress information from the original $150 \times 150 \times 3$ pixel input into the more compact 2048-dimensional embedding, which will both allow it to memorize idiosyncratic class labels and—as a side effect—discover regularities that support closure. By this argument, supervised training on true labels further boosts the network’s ability to extract structure meaningfully related to class identity. This structure allows the network to generalize to new images as well as to further support closure.

We chose to eliminate associative structure by shuffling labels, but an alternative approach might be to train an unsupervised architecture that uses only the input images, e.g., an autoencoder. We opted not to explore this alternative because label shuffling was a more direct and well-controlled manipulation; it allows us to re-use the base model architecture as is.

Replication on simpler architecture

To examine the robustness of the results we’ve presented thus far, we conducted a series of simulations with a smaller, simpler architecture. This architecture has three output classes, chosen randomly from the ImageNet data set, and three layers, each consisting of a convolutional mapping followed by max pooling. We train 8-10 networks with the same architecture and different weight initializations.

Figures 6e-6h show closure results for the simple architecture that correspond to the results from the larger architecture in Figures 6a-6d. This simple architecture produces the same pattern of closure effects as the larger Inception model, suggesting that closure is robust to architecture. Closure also appears to be robust to stimulus image diversity: Inception is trained on images from 1000 distinct classes; the simple net is trained on images from only 3 classes. However, we have observed lower bounds on the required diversity: When we train either model on one example per class, closure is not obtained.

The role of convolutional operators and local connectivity

Deep networks have become successful in vision tasks due to adopting some basic architectural features of the mammalian visual system, specifically, the assumptions of local connectivity and equivariance (Fukushima et al., 1983). Local connectivity in a topographic map indicates that a detector in one region of the visual field receives input only from its local neighborhood in the visual

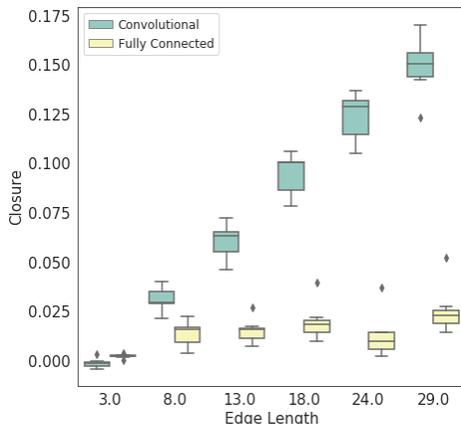


Figure 7: Exploring closure on convolutional versus fully connected architectures. Only the convolutional net achieves a closure effect, as indicated by the nonzero slope of the edge length vs. closure function.

field. Equivariance indicates that when presented with the same local input, detectors in different parts of the topographic map respond similarly. These properties are attained in deep nets via convolutional architectures with weight constraints.

To evaluate the role of these architectural constraints, we compare a ConvNet with the generic alternative, a *fully connected* architecture (*FCNet*) with dense (non-local) connectivity and no built in equivariance. Because FCNets do not perform as well on complex vision tasks, we were unable to train an FCNet on the full ImageNet data set to achieve performance comparable to our baseline ConvNet. Without matching the two architectures on performance, any comparison confounds architecture and performance. Consequently, we trained small instances of both architectures on just three randomly selected classes from ImageNet, allowing us to match the ConvNet and FCNet on performance. We replicated this simulation 7 times for robustness.

Figure 7 compares the closure effect for ConvNets and FCNets. The penultimate layer of representation is used to assess closure for both architectures. While the ConvNet evidences closure, the FCNet does not. Taking this finding together with the fact that the untrained ConvNet exhibits some degree of closure (Figures 6c and 6g), we infer that some aspect of the ConvNet structure facilitates the induction of closure.

Levels of representation and closure

Thus far, we have investigated closure at the penultimate layer of a network, on the assumption that this representation would be the most abstract and therefore most likely to encode object shape. However the deep Inception architecture we tested has 16 major layers, some of which involve multiple convolutional transformation and pooling operations. A priori, observing closure in early layers seems unlikely because the receptive fields of neurons in these layers have spatially constrained receptive fields, and closure requires the registration of Gestalt properties of the shapes. (Our test of closure will not false trigger based on local visual edge similarity because we compare images with distinct θ_{global} .)

In Figure 8a, we show the closure effect for representations in the last eleven layers of Inception. ‘Mixed_7c’ is the layer whose representation we have previously reported on. The graph legend is ordered top to bottom from shallowest to deepest layer. While all of the eleven layers show closure, closure is weaker for the shallower layers, labeled ‘Mixed_5’, than the deeper layers, labeled ‘Mixed_6’ and ‘Mixed_7’. We do not have a definitive explanation for why the effect is slightly weaker in the deeper ‘Mixed_7’ layers than in the shallower ‘Mixed_6’ layers, though we suspect it is a consequence of training the model for classification. Classification does not require any information other than class labels to be transmitted through the network. Consequently, the net is not encouraged to preserve a shape representation through all layers, and the net possibly discards irrelevant shape information in order to optimize inter-class discrimination.

In Figure 8b, we depict the closure curves for layers of the simple net, from the shallowest hidden layer, ‘conv2d_1’, to the penultimate layer, ‘fc_finale’. For this architecture, only the penultimate

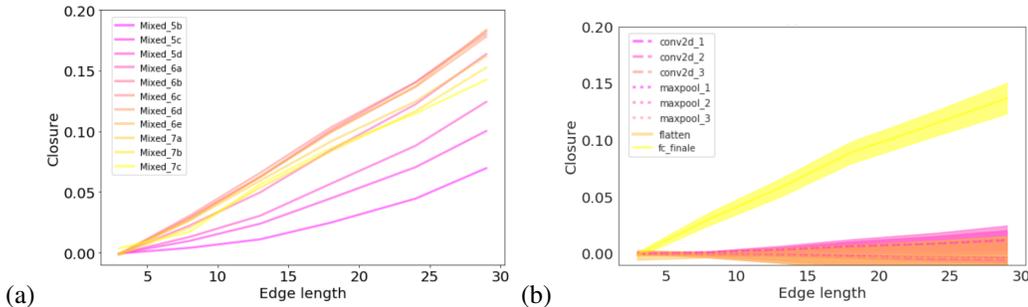


Figure 8: (a) The closure effect for the final eleven layers of the Inception architecture. Previously, we assessed closure only at layer ‘Mixed_7c’, but the lower layers also show varying degrees of closure. (b) The closure effect for each layer of the small ConvNet. Previous results have read out from the ‘fc_finale’ layer. In both graphs, variability over images in the strength of closure is shown with uncertainty shading.

layer shows closure. In the penultimate layer, each neuron can respond to information anywhere in the visual field.

Consistent across Inception and the simple net, representations at the shallow layers are not sufficiently abstract to encode Gestalt closure. This follows from the local feedforward connectivity of the architectures and gradual collapsing (pooling) of information across increasingly wider receptive fields at deeper stages of both architectures.

Although filling-in phenomena are observed in early stages of visual cortex (von der Heydt et al., 1984), it's possible that these effects are not causally related to Gestalt perception or are due to feedback projections, which our models lack. But our results are otherwise consistent with the view that lower stages of neural nets capture spatially local, low-order statistics whereas higher stages capture spatially global, high-order statistics (Bau et al., 2017; Mozer, 1991).

Discussion

Our work follows a long tradition in computational modeling of using neural network models to explain qualitative aspects of human perception (e.g., Rumelhart et al., 1988; Mozer, 1991). The strength of computational methods over behavioral or neuroscientific methods as an investigative tool is that models can be precisely manipulated to determine the specific model properties and inputs that are necessary and sufficient to explain a phenomenon. Further, we can do more than merely observe a model's input-output behavior; we can probe its internal representations and directly determine what it is computing.

We began with the conjecture that Gestalt laws need not be considered as primitive assumptions underlying perception, but rather, that the laws themselves may arise from a more fundamental principle: adaptation to statistics of the environment. We sought support for this conjecture through the detailed study of a key Gestalt phenomenon, closure. Using a state-of-the-art deep neural network model that was pretrained to classify images, we showed that in the model:

- *Closure depends on natural image statistics.* Closure is obtained for large neural networks trained to classify objects, and even for a smaller net trained to discriminate only a few object classes, but it is not obtained when a net is trained on white noise images or shuffled-pixel images. While shuffled-pixels have the same statistics as natural images, networks with local receptive fields are unable to extract spatially local structure due to the fact that the pixel neighborhood has been randomly dispersed in a shuffled image.
- *Closure depends on the architecture of convolutional networks.* The extraction of image regularities is facilitated by two properties of ConvNets: spatially local receptive fields and equivariance. Fully connected networks, which lack these forms of inductive bias, do not obtain closure. The inductive biases are sufficiently strong that even an untrained ConvNet obtains a weak degree of closure.
- *Closure depends on learning to categorize the world in a meaningful way.* Networks trained to associate images with their correct categorical label produce much larger closure effects than networks trained to associate images with random labels. In the former case, the labels offer a clue about what features should be learned to systematically discriminate categories (Lupyan, 2012). In the latter case, the labels misdirect the net to discover features that, by chance, happen to be present in a random collection of images that were assigned the same label.

Our simulation experiments suggest that these dependencies are both necessary and sufficient for a computational system to produce closure. The system need not have innate sensitivity to closure, nor does it need to *learn* closure per se. Rather, closure emerges as a byproduct of learning to represent real-world objects and categories.

Although we made this argument specifically for closure, the same argument applies to other Gestalt laws that have underpinnings in natural scene statistics (Brunswik and Kamiya, 1953; Elder and Goldberg, 2002; Geisler et al., 2001; Krüger, 1998; Sigman et al., 2001). Given structure in the environment and the existence of powerful learning architectures and mechanisms, one need not consider the Gestalt laws as primitives. Rather, the Gestalt laws can emerge via adaptation to the statistical structure of the environment.

One limitation of our work is that we do not claim our trained neural net classifier is a cognitive model, i.e., a model of the cognitive processes involved in biological perception. At some very coarse level of analysis—e.g., the fact that the model learns from experience to classify naturalistic images—the model does replicate human abilities. However, we have not tried to explain detailed and specific experimental phenomena in the way one typically does in cognitive modeling. A phenomenon we most likely cannot explain in our model is human performance in the classification image paradigm—the paradigm discussed earlier in which noisy stimuli are averaged together to infer illusory contours. A pretrained neural net much like the one we studied does not show the same effect as human subjects (Baker et al., 2018). However, the model is not inconsistent with the phenomenon; rather it would have to be expanded to explain a full range of phenomena. In this case, the expansion might involve a different input representation, such as edge detectors, or feedback projections to achieve better biological realism (Spoerer et al., 2017).

We have focused our study on modal completion, which involves the detection of illusory shapes. Amodal completion, which involves the detection of occluded shapes, provides a further opportunity to evaluate the consequences of learning natural scene statistics. There is some debate whether modal and amodal completion rely on the same visual mechanisms in humans (Anderson, 2007; Kellman et al., 1998). In our model, modal completion likely arises because explicit perceptual evidence about the edges of an object in a natural scene is occasionally absent; thus, natural image perception demands inferring the missing features. In our model, amodal completion would arise for the analogous reason: the presence of partially occluded objects where missing edge features again must be inferred. The case for occlusion in natural scenes is much stronger than the case for missing features, and thus an adaptation-based account of amodal completion seems quite natural. Indeed, the Zemel et al. (2002) study (Figure 2) provides clear evidence that completion can be learned under occlusion as a consequence of experience.

In the late 1980s, the Connectionist movement focused on the role of learning in perception and cognition, demonstrating that abilities that one might previously have thought to have been built into a cognitive system could emerge as a consequence of simple learning rules. Most connectionist architectures were fairly generic—typically fully connected neural networks with one hidden layer. While such architectures showed promise, it was far from clear that these architectures could scale up to human-level cognitive abilities. Modern deep learning architectures have clearly scaled in a manner most did not imagine in the 1980s. Large language models show subtle linguistic skills and can express surprisingly broad knowledge about the world (Raffel et al., 2019); large vision models arguably match or surpass human abilities to label images (Xie et al., 2019). Our work suggests that in the modern era, the combination of a sophisticated neural architecture (e.g., a ConvNet) and scaling the size of models may be sufficient to broaden the range of cognitive phenomena that are emergent from more basic principles of learning and adaptation. Our work bridges the gap between analyses indicating that perceptual mechanisms are consistent with natural scene statistics (Burge et al., 2010; Brunswik and Kamiya, 1953) and claims that statistical learning is essential to understanding human information processing (Frost et al., 2019). The synthesis leads to a view of the human perceptual system that is even more elegant than the Gestaltists imagined: a single principle—adaptation to the statistical structure of the environment—might suffice as fundamental.

Methods

Models

In our large simulation experiments, we leverage a state-of-the-art, pretrained image classification network, *Inception*, trained on the 1000-class ImageNet data set (Szegedy et al., 2016). Input to this model is a 150×150 pixel color (RGB) image, and output a 1001-dimensional activation vector whose elements indicate the probability that the image contains the corresponding object class. (The ImageNet task involves 1000 classes; the additional class is a ‘none of the above’ category.) *Inception* has been trained on 1.2 million images from the ImageNet data set (Deng et al., 2009) to assign each image to one of a thousand object classes. We train with standard data augmentation methods including: horizontal flips, feature-wise normalization, aspect ratio adjustment, shifts, and color distortion. In most simulations, we read out representations from the penultimate layer of the net, known as *Mixed_7c*, which consists of a 2048-dimensional flat vector. The penultimate layer of a deep net is commonly assumed to embody a semantic representation of the domain (visual objects). For example, in transfer learning and few-shot learning models, this layer is used for encoding novel

object classes (e.g., Scott et al., 2018; Yosinski et al., 2014). One of our experiments reads out from earlier layers of the network.

We also explore a *simple convolutional* architecture consisting of three pairs of alternating convolutional and max pooling layers followed by a fully connected layer and a single output unit trained to discriminate between three classes, randomly chosen from ImageNet. A *fully connected* variant of the architecture replaces the convolutional and pooling blocks with fully connected layers. For these simple models, the embedding is the penultimate layer of the network.

For the sanity-check experiments (CD and BD models), we used the simple convolutional architecture with the three output classes with a single output which performs a binary discrimination (disordered versus complete and aligned for CD; black backgrounds versus white backgrounds for BD). The CD and BD models are trained on 75% of the 768 distinct complete-aligned-disordered triples; the remainder form a validation set, which reaches 100% accuracy and is used for evaluating the model. Five replications of the CD and BD models are trained with different random initial seeds to ensure reliability of results.

Further details on all models are provided in the Supplementary Materials.

Stimuli

We compare three critical conditions (Figure 3c-e): complete triangles, triangle fragments with aligned corners, and fragments with disordered corners. Each stimulus is rendered in a 150×150 pixel image and the Euclidean distance between vertices is 116 pixels. Rather than testing models with more elaborate images (e.g., Figures 3a,b), we chose to use the simplest images possible that could evoke closure effects, for two reasons. First, with more complex and naturalistic images, we found that it was difficult to control for various potential confounds. Second, complex examples like the Kanizsa triangle potentially involve both modal completion (the white triangle in the foreground of Figure 3a) and amodal completion (the outline triangle and the solid black disks behind the foreground triangle) completion). We wished to simplify and focus specifically on the Gestalt principle of closure.

We manipulated various properties of the stimuli, as depicted in Figure 4. For all conditions, the stimuli varied in the global orientation of the triangle or fragments, which we refer to as θ_{global} , the *background* (light on dark versus dark on light), and the *position* of the object center in the image. For the disordered condition, we varied the orientation of the corners with respect the orientation of corners in the aligned-fragment condition, which we refer to as θ_{local} . And finally, for the disordered and aligned conditions, we varied the length of the edges extending from the fragment corners, which we refer to as *edge length*.

Edge length is centrally related to the phenomenon of interest. Edge length, or equivalently, the gap between corners, influences the perception of closure, with smaller gaps leading to stronger closure (Elder and Zucker, 1993; Jakel et al., 2016). The remaining properties—background color, local and global orientation, and image position—are manipulated to demonstrate invariance to these properties. If sensitivity to any of these properties is observed, one would be suspicious of the generality of results. Further, these properties must be varied in order to avoid a critical confound: the complete image (Figure 3d) shares more pixel overlap with the aligned fragments (Figure 3c) than with the disorderd fragments (Figure 3d). We therefore must ensure that any similarity of response between complete and aligned images is not due to pixel overlap. We accomplished this aim by always comparing the response to complete and fragment images that have different θ_{global} and different image positions. However, when comparing representations, we always match the images in background color because neural nets tend to show larger magnitude responses to brighter images.

The background color has two levels, black and white. The position is varied such that the stimuli could be centered on the middle of the image or offset by -8 pixels from the center in both x and y directions, resulting in two distinct object locations. The global orientation is chosen from eight equally spaced values from 0° to 105° . (Symmetries make additional angles unnecessary. A 120° triangle is identical to a 0° triangle.) The local orientation of the disordered corners is rotated from the aligned orientation by 72° , 144° , 216° , or 288° . The edge length is characterized by the length of an edge emanating from the vertex; we explored six lengths: 3, 8, 13, 18, 24, and 29 pixels, which corresponds to removal of between 95% and 50% of the side of a complete triangle to form an aligned image. These manipulations result in $2 \times 2 \times 8 = 32$ complete triangles, $2 \times 2 \times 8 \times 6 = 192$ aligned fragments, and $2 \times 2 \times 8 \times 6 \times 4 = 768$ disordered fragments, totalling 992 distinct stimulus images.

Quantitative measure of closure

Elder and Zucker (1993) studied closure via a visual-search task that required human participants to discriminate among simple shapes like those in Figure 4. Although we could, in principle, ask a network to perform this task, additional assumptions would be required about response formation and initiation. Instead, we claim simply that the difficulty of the task depends on the similarity of internal representations: if {complete, aligned} pairs are more similar to one another than {complete, disordered} pairs, discrimination ability and response latency should be longer by any sensible biologically-plausible read out process (e.g., Ratcliff and McKoon, 2008). This claim allows us to focus entirely on the representations themselves, via a quantitative measure of closure:

$$C_i = s(f(\mathbf{a}_i), f(\mathbf{c}_i)) - s(f(\mathbf{d}_i), f(\mathbf{c}_i)),$$

where i is an index over matched image triples consisting of a complete triangle (\mathbf{c}_i), aligned fragments (\mathbf{a}_i), and disordered fragments (\mathbf{d}_i); $f(\cdot) \in \mathbb{R}^m$ is the neural net mapping from an input image in $\mathbb{R}^{150 \times 150}$ to an m -dimensional embedding, and $s(\cdot, \cdot)$ is a similarity function (Figure 4). Consistent with activation dynamics in networks, we use a standard similarity measure, the cosine of the angle between the two vectors,²

$$s(\mathbf{x}, \mathbf{y}) = \frac{f(\mathbf{x})f(\mathbf{y})^T}{|f(\mathbf{x})| |f(\mathbf{y})|}.$$

The triples are selected such that \mathbf{a}_i and \mathbf{d}_i are matched in θ_{global} position, both differ from \mathbf{c}_i in θ_{global} , and all three images have the same background color (black or white). These constraints ensure that there is no more pixel overlap (i.e., Euclidean distance in image space) between complete and aligned images than between complete and disordered images.

We test 768 triples by systematically pairing each of the 768 distinct disordered images with randomly selected aligned and complete images, subject to the constraints in the previous paragraph. Each of the 192 aligned images in the data set is repeated four times, and each of the 32 complete images is repeated 24 times.

We compute the mean closure across triples, $\bar{C} \in [-1, +1]$. This measure is $+1$ when the complete image yields a representation identical to that of the aligned image and orthogonal to that of the disordered image. These conditions are an unambiguous indication of closure because the model cannot distinguish the complete triangle from the aligned fragments. Mean closure \bar{C} is 0 if the complete image is no more similar to the aligned than disordered images, contrary to what one would expect by the operation of Gestalt grouping processes that operate based on the alignment of fragments to construct a coherent percept similar to that of the complete triangle. Mean closure \bar{C} may in principle be negative, but we do not observe these values in practice.

Although our measure of representational similarity is common in the deep learning literature, the neuroimaging literature has suggested other notions of similarity, e.g., canonical correlation analysis (Hårdle and Simar, 2007) and representational similarity analysis (Kriegeskorte et al., 2008). However, these measures are primarily designed to compare signals of different dimensionality (e.g., brain-activity measurement and behavioral measurement).

Acknowledgements

We are grateful to Mary Peterson for insightful feedback on an earlier draft of the paper, and to Bill Freeman and Ruth Rosenholtz for helpful discussions about the research. Special thanks to Corbin Cunningham for his advice on our experiment designs.

References

- Anderson, B. L. (2007). Filling-in models of completion: Rejoinder to kellman, garrigan, shipley, and keane (2007) and albert (2007). *Psychological Review*, 114:509–525.
- Ba, J. and Caruana, R. (2014). Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, pages 2654–2662.

²We assume $s(\mathbf{x}, \mathbf{y}) = 0$ if both $|f(\mathbf{x})| = 0$ and $|f(\mathbf{y})| = 0$.

- Baker, N., Kellman, P. J., Erlikhman, G., and Lu, H. (2018). Deep convolutional networks do not perceive illusory contours. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 1310–1315, Austin, TX. Cognitive Science Society.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*.
- Behrens, R. R. (1998). Art, design and gestalt theory. *Leonardo*, 31(4):299–303.
- Bender, L. (1938). A visual motor gestalt test and its clinical use. *Research Monographs, American Orthopsychiatric Association*.
- Brunswik, E. and Kamiya, J. (1953). Ecological cue-validity of 'proximity' and of other gestalt factors. *The American Journal of Psychology*, 66(1):20–32.
- Burge, J., Fowlkes, C. C., and Banks, M. S. (2010). Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *Journal of Neuroscience*, 30:7269–7280.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.
- Desolneux, A., Moisan, L., and Morel, J.-M. (2007). *From gestalt theory to image analysis: a probabilistic approach*, volume 34. Springer Science & Business Media.
- Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, 113:501–517.
- Elder, J. and Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research*, 33(7):981–991.
- Elder, J. H. and Goldberg, R. M. (2002). Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):324–353.
- Frost, R., Armstrong, B. C., and Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 145:1128–1153.
- Fukushima, K., Miyake, S., and Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):826–834.
- Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41:711–724.
- Gold, J. M., Murray, R. F., Bennett, P. J., and Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, 10:663–666.
- Grossberg, S. (2014). How visual illusions illuminate complementary brain processes: illusory depth from brightness and apparent motion of illusory contours. *Frontiers in Human Neuroscience*, 8:854–866.
- Härdle, W. and Simar, L. (2007). *Applied multivariate statistical analysis*, volume 22007. Springer.
- Holmes, D. S. (1968). Search for "closure" in a visually perceived pattern. *Psychological Bulletin*, 70(5):296–312.
- Jakel, F., Singh, M., Wichmann, F. A., and Herzog, M. H. (2016). An overview of quantitative approaches in gestalt perception. *Vision Research*, 126:3 – 8. Quantitative Approaches in Gestalt Perception.
- Kalar, D. J., Garrigan, P., Wickens, T. D., Hilger, J. D., and Kellman, P. J. (2010). A unified model of illusory and occluded contour interpolation. *Vision Research*, 50:284–299.
- Kellman, P. J., Yin, C., and Shipley, T. F. (1998). A common mechanism for illusory and occluded object completion. *Journal of Experimental Psychology: Human Perception and Performance*, 24:859–869.

- Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: a critical review. *Psychological bulletin*, 112(1):24.
- Kimchi, R., Yeshurun, Y., Spehar, B., and Pirkner, Y. (2016). Perceptual organization, visual attention, and objecthood. *Vision Research*, 126:34 – 51. Quantitative Approaches in Gestalt Perception.
- Kornblith, S., Shlens, J., and Le, Q. V. (2018). Do better ImageNet models transfer better? *CoRR*, abs/1805.08974.
- Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.
- Krüger, N. (1998). Collinearity and parallelism are statistically significant second-order relations of complex cell responses. *Neural Processing Letters*, 8:117–129.
- Lin, H. W., Tegmark, M., and Rolnick, D. (2017). Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247.
- Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3(54):1–13.
- Mozer, M. C. (1991). *The Perception of Multiple Objects: A Connectionist Approach*. MIT Press, Cambridge, MA, USA.
- Peterson, M. A. (2019). Past experience and meaning affect object detection: A hierarchical bayesian approach. In Federmeier, K. D. and Beck, D. M., editors, *Knowledge and Vision*, volume 70 of *Psychology of Learning and Motivation*, pages 223 – 257. Academic Press.
- Peterson, M. A. and Gibson, B. S. (1994). Must figure-ground organization precede object recognition? an assumption in peril. *Psychological Science*, 5(5):253–259.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.
- Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20:873–922.
- Ren, X. and Malik, J. (2003). Learning a classification model for segmentation. In *null*, page 10. IEEE.
- Rensink, R. A. and Enns, J. T. (1998). Early completion of occluded objects. *Vision Research*, 38:2489–2505.
- Ringach, D. L. and Shapeley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Research*, 36(19):3037–3050.
- Ringach, D. L. and Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision research*, 36(19):3037–3050.
- Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.
- Sanguinetti, J. L., Trujillo, L. T., Schnyer, D. M., Allen, J. J., and Peterson, M. A. (2016). Increased alpha band activity indexes inhibitory competition across a border during figure assignment. *Vision Research*, 126:120–130.
- Schultz, D. P. and Schultz, S. E. (2015). *A history of modern psychology*. Cengage Learning.
- Scott, T. R., Ridgeway, K., and Mozer, M. C. (2018). Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’ 18, page 76–85, Red Hook, NY, USA. Curran Associates Inc.

- Sigman, M., Cecchi, G. A., Gilbert, C. D., and Magnasco, M. O. (2001). On a common circle: natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences*, 98:1935–1940.
- Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, 8:1551–1564.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.
- Todorovic, D. (2008). Gestalt principles. *Scholarpedia*, 3(12):5345.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454.
- von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, 224(4654):1260–1262.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., and von der Heydt, R. (2012a). A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172.
- Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., and van Leeuwen, C. (2012b). A century of gestalt psychology in visual perception: II. conceptual and theoretical foundations. *Psychological bulletin*, 138(6):1218.
- Wertheimer, M. (1923). Laws of organization in perceptual forms. *A source book of Gestalt Psychology*.
- Westheimer, G. (1999). Gestalt theory reconfigured: Max wertheimer’s anticipation of recent developments in visual neuroscience. *Perception*, 28(1):5–15.
- Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2019). Self-training with noisy student improves imagenet classification.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, page 3320–3328, Cambridge, MA, USA. MIT Press.
- Zemel, R. S., Behrmann, M., Mozer, M. C., and Bavelier, D. (2002). Experience-dependent perceptual grouping and object-based attention. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):202–217.
- Zhang, C., Bengio, S., Hardt, M., Mozer, M. C., and Singer, Y. (2020). Identity crisis: Memorization and generalization under extreme overparameterization. In *International Conference on Learning Representations*.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530.
- Zinker, J. (1977). *Creative process in Gestalt therapy*. Brunner/Mazel.

Supplementary Materials

Experimental setup details

Simple network Each network has $n_c \in \{3, 6, 9\}$ classes (randomly chosen from Imagenet dataset) with $n_l \in \{3, 5, 7\}$ layers. These layers are either convolutional or fully connected. For convolutional networks (ConvNets), we iterate between convolutional and max-pooling layers n_l times to predict n_c classes. In each layer, the number of neurons increases by 16, except the final layer always has 512 neurons regardless of the n_c . For fully connected networks (FCNets), we first flatten the input image, then add n_l fully-connected layers. All networks are learned using the RMSProp method with 0.001 (for convolutional) 0.0001 (for FCNet, smaller rate was critical for learning) learning rate for 100 epochs. The training dataset was prepared with standard data augmentation: feature-wise normalization, linear translation (0.02 range) and horizontal flips.

Sanity check network This network has the identical setup to the simple network above with 3 convolution layers ($n_l = 3$) and 2 classes ($n_c = 2$).

More complex network (Inception) This network uses more complex and more widely used InceptionV3 network architecture (Szegedy et al., 2016). This network was trained on 1.2 million ImageNet images, with similar augmentation to the simple network: horizontal flips, featurewise normalization, aspect ratio adjustment, shifts, and color distortion. It was trained to top-5 accuracy of 92% over 120 epochs with a batch size of 4096. The learning rate and weights decay followed those of Kornblith et al. (2018). A depiction of the architecture is shown on https://microscope.openai.com/models/inceptionv3_slim.

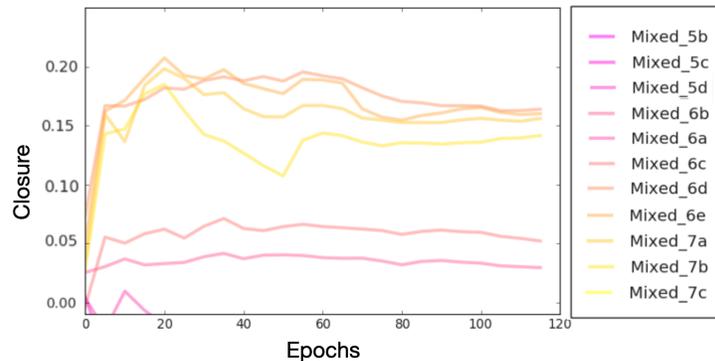


Figure 9: The closure effect in the last layer reaches its peak earlier in the training process, then decreases somewhat as it converges. Other layers seem to continue to fluctuate before converging. For degenerated networks, the effect is quickly forgotten. Showing results from Inception.

The closure effect during training before convergence

We hypothesized that the closure effect would increase during training and will converge, similar to a typical validation accuracy curve. This varies depending on the layer. The closure effect reaches its peak earlier in the iteration, then fluctuates as it learns, then forgets the effect slightly as it converges in both simple network and Inception (Fig. 9). This is typically observed in higher layers (e.g., Mixed 7a and above in Inception). In lower layers (e.g., Mixed 6d and lower), the closure effect increases then converges.

On the other hand, networks trained with degenerate training data (e.g., shuffled pixels, shuffled labels) start with some closure effect, since untrained network exhibits some closure effect. However the closure effect drops immediately and stays close to zero in the duration of training.

The rapid decrease of the closure effect in networks other than the baseline network aligns with our findings; in the process of trying to fit to random data, the network loses some of the initial feature extraction properties that it had had due to convolutional operations (Ulyanov et al., 2018), and the closure effect is also lost.

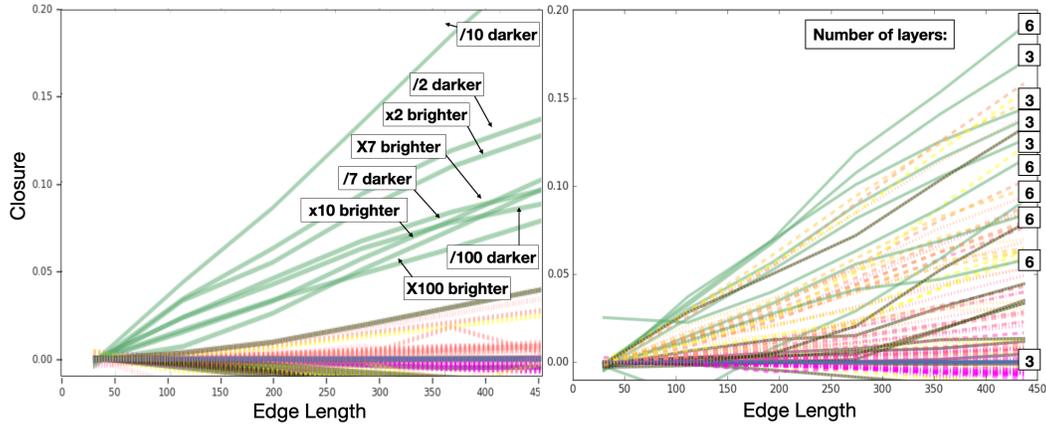


Figure 10: Varying brightness of stimuli and the number of layers: The closure effect does not have clear pattern of change as images become brighter/darker or as the depth of network and/or number of classes trained change.

The fluctuation during learning before convergence is an interesting symptom, and may benefit from further study. This hints that the closure effect may reflect a prediction-related signal that can be useful (e.g., to determine stopping conditions).

The closure effect is uninfluenced simple input manipulations

In this experiment, we try to invalidate a hypothesis that a seemingly meaningless input manipulation on the training data (e.g., image brightness) will arbitrarily influence the closure effect. For example, we should not be able to increase/decrease the closure effect by simply making training images brighter/darker (i.e., multiplying/dividing images with constant).

There is no strong pattern between closure effects and brightness of the training images (Fig. 10). Note that the variance in each run reflects the amount of information lost by multiplying or dividing each pixel and saturating them (e.g., multiplying images with 10 cause some images to be no longer identifiable). Naturally in conditions with no strong closure effect (e.g., shuffled pixels), there was no difference of the effect from brightness variation.

The closure effect as number of classes and number of layers vary.

We discovered that there seems to be no strong correlation between the depth of the network and the closure effect (Figures 11). Given that the optimal choice of the depth of a network is an unsolved problem, despite much work Lin et al. (2017); Ba and Caruana (2014), and it is not always true that adding more layers improve the model's performance Ba and Caruana (2014), it is reasonable that we cannot influence the closure effect by simply adding more layers.

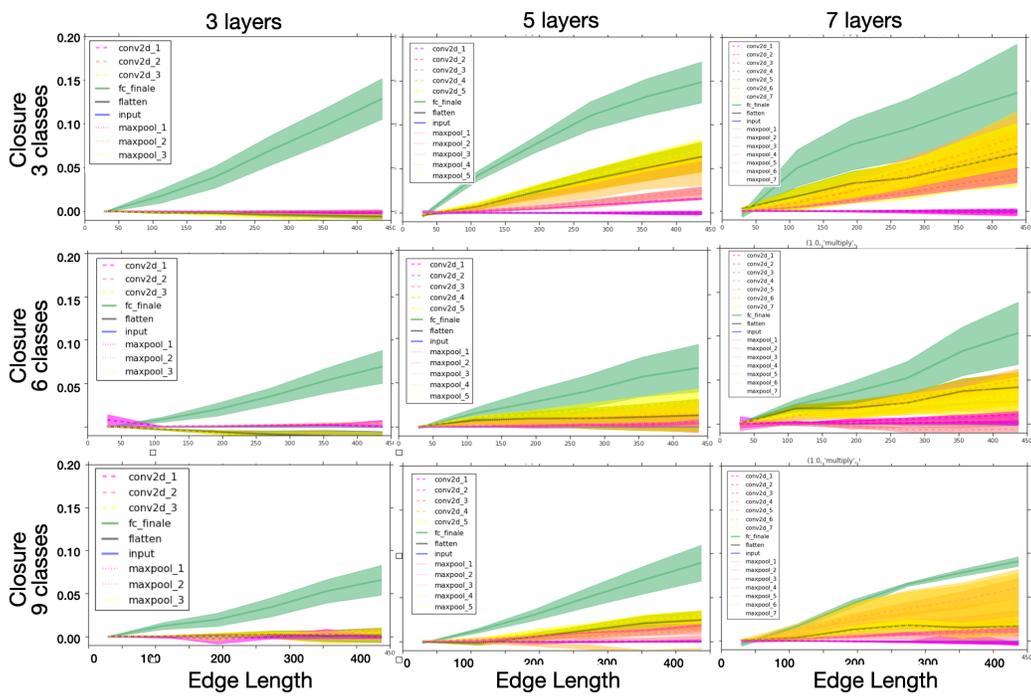


Figure 11: The closure effect for while changing number of classes, nc , and number of layers, nl . The effect does not seem to have clear pattern with respect to these factors.