

<b>Noname manuscript No.</b> (will be inserted by the editor)
--

# Deep Nets: What have they ever done for Vision?

Alan L. Yuille · Chenxi Liu

Received: date / Accepted: date

**Abstract** This is an opinion paper about the strengths and weaknesses of Deep Nets for vision. They are at the center of recent progress on artificial intelligence and are of growing importance in cognitive science and neuroscience. They have enormous successes but also clear limitations. There is also only partial understanding of their inner workings. It seems unlikely that Deep Nets in their current form will be the best long-term solution either for building general purpose intelligent machines or for understanding the mind/brain, but it is likely that many aspects of them will remain. At present Deep Nets do very well on specific types of visual tasks and on specific benchmarked datasets. But Deep Nets are much less general purpose, flexible, and adaptive than the human visual system. Moreover, methods like Deep Nets may run into fundamental difficulties when faced with the enormous complexity of natural images which can lead to a combinatorial explosion. To illustrate our main points, while keeping the references small, this paper is slightly biased towards work from our group.

**Keywords** Deep Neural Networks · Computer Vision · Success · Limitation · Cognitive Science · Neuroscience

## 1 Introduction

In the last few years Deep Nets have enabled enormous advances in computer vision and the study of

---

For those readers unfamiliar with Monty Python see: <https://youtu.be/Qc7HmhrgTuQ>

---

Alan L. Yuille  
Johns Hopkins University, Baltimore, MD, USA  
E-mail: alan.l.yuille@gmail.com

Chenxi Liu  
Johns Hopkins University, Baltimore, MD, USA  
E-mail: cxliu@jhu.edu

biological visual systems. But as researchers in these areas, we find ourselves having mixed feelings about them. On the one hand, we marvel at their successes and how they have led to amazing results on some real world tasks and, in academic settings, their performance on benchmarked datasets almost always outperforms alternative approaches. But, on the other hand, we are aware of their limitations and concern about the hype that surrounds them. Several recent papers (Darwiche, 2018; Marcus, 2018) have critiqued Deep Nets from the perspectives of machine reasoning and cognitive science, arguing that though Deep Nets are useful as a tool they will need to be combined with alternative approaches in order to achieve human level intelligence. The nature of our research means that we interact with research faculty in many disciplines (cognitive science, computer science, applied mathematics, engineering, neuroscience, physics, and radiology) and the Deep Nets are a frequent topic of conversation. We find ourselves spending half the time criticizing Deep Nets for their limitations and the other half praising them and defending them against their critics (not infrequently we are confidently told that “Deep Nets can never do xxx” when we already know that they can). This opinion paper attempts to provide a balanced viewpoint on the strengths and weaknesses of Deep Nets for studying vision.

The organization of this article is as follows. In Section 2 we discuss the history of neural networks and its tendency to boom and bust. Section 3 describes a few of the successes of Deep Nets while also mentioning the caveats and fine print. In Section 4 we discuss the limited understanding of the internal workings of Deep Nets. Section 5 surveys their potential for helping to construct theories of biological visual systems, but also their limited relationships to real neurons and neu-

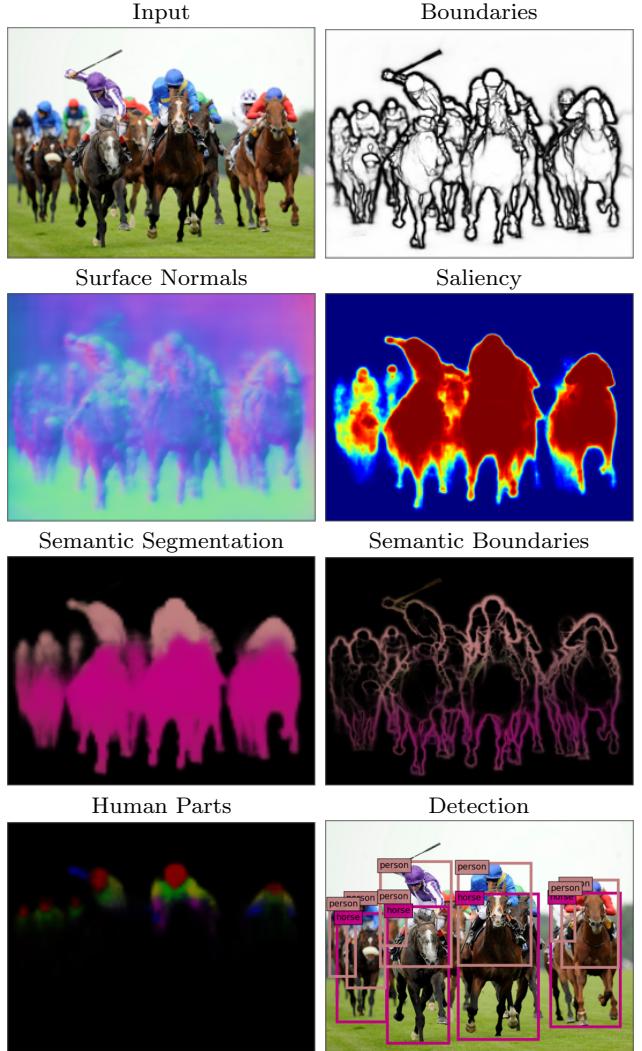
ral circuits. In Section 6 we discuss the challenges that Deep Nets are now grappling with. Section 7 is more speculative and argues that as vision researchers attempt to model increasingly complex visual tasks they will face a combinatorial explosion which Deep Nets may be unable to handle.

## 2 Some History

We are in the third wave of neural network approaches. The first two waves — 1950s–1960s and 1980s–1990s — generated considerable excitement but slowly ran out of steam. Despite a few exceptions, the overall performance of neural networks was disappointing for machines (artificial intelligence/machine learning) and for understanding biological vision systems (neuroscience, cognitive science, psychology). But the third wave — 2000s–present — is distinguished because of the dramatic success of Deep Nets on many large benchmarked problems and their industrial application to real world tasks. It should be acknowledged that almost all the basic ideas of many of the currently successful neural networks were developed during the second wave. But their strengths were not appreciated until the availability of big datasets and the ubiquity of powerful computers (e.g., GPUs) which only became available after 2000 and which fueled the third wave.

The rise and falls of these neural network waves reflect changes in intellectual fashion and the varying popularity of other approaches. The second wave of neural networks was partly driven by the perceived limitations of classic artificial intelligence where disappointing results and accusations of over-promising led to an AI winter in the mid-1980s. In turn, the decline of the second wave corresponded to the rise of support vector machines, kernel methods, and related approaches. Credit is due to those neural network researchers who carried on despite discouragement through the troughs of the waves when it was sometimes hard to publish neural network papers. The pendulum has now swung again and it sometimes seems hard to publish anything that is not neural network related. We suspect that progress would be faster if researchers resisted the attraction of fashions and instead pursued a diversity of approaches and techniques. It is also worrying that the courses for students often tends to follow the latest fashions and ignore the older techniques (until they are rediscovered).

The current successes of neural networks are mainly for artificial intelligence tasks where they have made big advances in tasks like face recognition (now working on datasets of tens of millions of people) and on medical image analysis. Neural networks are increasingly being



**Fig. 1** Figure taken from Kokkinos (2017). A wide variety of vision tasks can be performed by Deep Nets. These include: boundary detection, semantic segmentation, semantic boundaries, surface normals, saliency, human parts, and object detection.

used to model the mind and brain but their relations to real neurons and neural circuits should be treated with caution. Although artificial neural networks were inspired by biology it must be acknowledged that real neurons are much more complex and understanding real neural circuits remains one of the most fundamental challenges of neuroscience.

## 3 The Successes, with the Fine Print

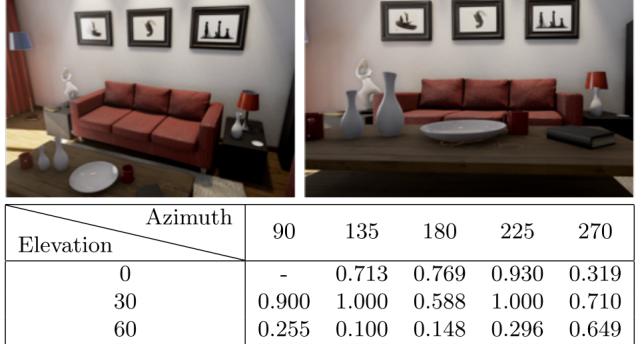
The computer vision community was fairly skeptical about Deep Nets until the impressive performance of AlexNet (Krizhevsky et al., 2012) for classifying objects in ImageNet (Deng et al., 2009). This classification

task assumes there is a foreground object which is surrounded by a limited background region, so the input is similar to one of the red boxes of the bottom right image in Figure 1. AlexNet’s success stimulated the vision community leading to a variety of Deep Net architectures with increasingly better performance on object classification, e.g., (Simonyan and Zisserman, 2015; He et al., 2016; Liu et al., 2018).

Deep Nets were also rapidly adapted to other visual tasks such as object detection, where the image contains one or more objects and the background is much larger, e.g., the PASCAL challenge (Everingham et al., 2010). For this task, Deep Nets were augmented by an initial stage which made proposals for possible positions and sizes of the objects and then applied Deep Nets to classify the proposals (current methods train the proposals and objects together in what is called “end-to-end”). These methods outperformed the previous best methods, the Deformable Part Models (Felzenszwalb et al., 2010), for the PASCAL object detection challenge (PASCAL was the main object detection and classification challenge before ImageNet). Other Deep Net architectures also gave enormous performance jumps in other classic tasks like edge detection, semantic segmentation, occlusion detection (edge detection with border-ownership), symmetry axis detection. Major increases also happened for human joint detection, human segmentation, binocular stereo, 3D depth estimation from single images, and scene classification. Several of these tasks are illustrated in Figure 1.

But although Deep Nets are very effective, almost always outperforming alternative techniques, they are not general purpose and their successes come with the following three restrictions.

Firstly, Deep Nets are designed for specific visual tasks. Most Deep Nets are designed for single tasks and a Deep Net designed for one task will not be well-suited for another. For example, a Deep Net designed for object classification on ImageNet cannot perform human parsing (i.e. the detection of human joints) on the Leeds Sports Dataset (LSD). There are, however, some exceptions and “transfer learning” sometimes makes it possible to adapt Deep Nets trained on one task to a closely related task provided annotated data is available for that task (see Section 6.1). Intuitively this happens because the features learned by the Deep Net captures image structures that are useful for both tasks. In addition, researchers have recently developed Deep Nets, e.g., UberNet (Kokkinos, 2017), which can perform up to four tasks with the same network. But, in general, there is a growing zoo of different Deep Net architectures designed for specific tasks which include cascades



**Fig. 2** Figure taken from Qiu and Yuille (2016). UnrealCV allows vision researchers to easily manipulate synthetic scenes, e.g. by changing the viewpoint of the sofa. We found that the Average Precision (AP) of Faster-RCNN (Ren et al., 2015) detection of the sofa varies from 0.1 to 1.0, showing extreme sensitivity to viewpoint. This is perhaps because the biases in the training cause Faster-RCNN to favor specific viewpoints.

of networks and supervision at several different levels of the network.

Secondly, Deep Nets which perform well on benchmarked datasets may fail badly on real world images outside the dataset. This is because the set of real world images is infinitely large and so it is hard for any dataset, no matter how big, to be representative of the complexity of the real world. This is an important issue which we will return to in Section 7. For now, we simply remark that all datasets have biases. These biases were particularly blatant in the early vision datasets and researchers rapidly learned to exploit them for example by exploiting the background context (e.g., detecting fish in Caltech101 was easy because they were the only objects whose backgrounds were water). Comparative studies showed that methods which performed well on some datasets often failed to generalize to others (Torralba and Efros, 2011). These problems are reduced, but still remain, despite the use of big datasets and Deep Nets. For example, background context remains problematic even for ImageNet (Zhu et al., 2017). Biases also occur if the dataset contain objects from limited viewing conditions, e.g., as shown in Figure 2, a Deep Net trained to detect sofas on ImageNet can fail to detect them if shown from viewpoints which were under-represented in the training dataset. In particular, Deep Nets are biased against “rare events” which occur infrequently in the datasets. But in real world applications, these biases are particularly problematic since they may correspond to situations where failures of a vision system can lead to terrible consequences, e.g., datasets used to train autonomous vehicles almost never contain babies sitting in the road. Similarly, datasets often tend to under-represent the hazardous factors which are

known to cause algorithm to fail, such as specularity for binocular stereo. We will return to this example in Section 6.3.

Thirdly, almost all Deep Nets require annotated data for training and testing. This has the effect of biasing vision researchers to work on those visual tasks for which annotation is easy. For example, annotation for object detection merely requires specifying a tight bounding box around an object. But for other vision tasks, such as detecting the joint of a human, annotation is much harder and for some tasks it is almost impossible. There are methods which reduce the need for supervision as discussed in Section 6.1, and there is also the possibility of using synthetic stimuli (generated by computer graphics engines) which enables groundtruth to be available for all visual tasks. But realistic synthetic stimuli are limited and the vision community is reluctant to rely on it until they become sufficiently realistic.

In summary, Deep Nets are a set of tools which are constantly being refined and developed according to the needs of specific visual tasks. They almost all rely on fully supervised data, with caveats we will discuss later, and their performance can fail to generalize to images outside the dataset they have been trained on. Dataset biases are particularly problematic for vision due to the infinite complexity of real world images, as we will discuss in Section 7.

#### 4 Towards Understanding Deep Nets

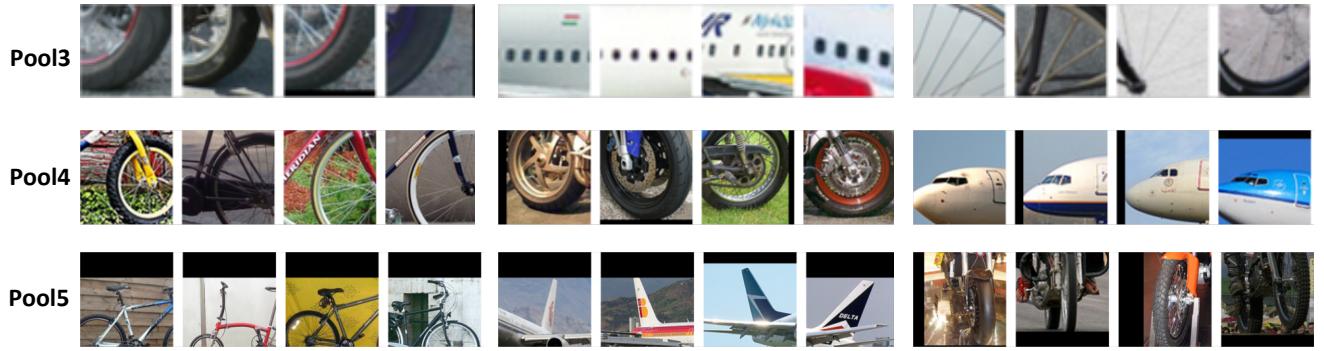
It is difficult to characterize what Deep Nets can do and to understand their inner workings. Theoretical results show that multi-layer perceptrons, and hence Deep Nets, can represent any input output function provided there are a sufficient number of hidden units (Hornik et al., 1989). But, as anybody who has proven theorems of this type is well aware (Xu et al., 1994), theoretical results which hold in the asymptotic limit are of limited utility. Much more valuable would be results which hold for limited numbers of hidden units and limited training data, but it is hard to see what meaningful theoretical results could be obtained for systems as complicated as Deep Nets.

At a more intuitive level it seems possible to get some rough understanding of Deep Nets at least when applied to visual tasks. The hierarchical structure of Deep Nets is similar to classical models of the visual cortex such as the NeoCognition (Fukushima and Miyake, 1982) and HMax (Riesenhuber and Poggio, 1999) and captures many of the intuitions which motivated these models. Deep Nets contain feature representations where those at lower levels have receptive fields of limited

sizes and which are sensitive to the precise positions of patterns. But as we ascend the hierarchy the receptive fields become larger and more sensitive to specific patterns, while being less concerned about their exact locations.

This can be partially understood by studying the activities of the internal filters/features of the convolutional levels of Deep Nets (Zeiler and Fergus, 2014; Yosinski et al., 2015). In particular, if Deep Nets are trained for scene classification then some convolutional layer filters roughly correspond to objects which appear frequently in the scene, while if the Deep Nets are trained for object detection, then some features roughly correspond to parts of the objects (Zhou et al., 2015). In detailed studies of a restricted subset of objects (e.g., vehicles), researchers (Wang et al., 2015) discovered regular patterns of activity of the feature vectors, called visual concepts, which corresponded approximately to the semantic parts of objects (with sensitivity to viewpoint), see Figure 3. But we acknowledge that while these studies are encouraging they remain fairly impressionistic and lack the precision of true understanding (e.g., these studies have not yet enabled researchers to learn models of objects and object-parts in an unsupervised manner).

This suggests the following rough conceptual picture of Deep Nets. The convolutional levels represent the manifold of intensity patterns at different levels of abstraction. The lowest levels represent local image patterns while the high levels represent larger patterns which are invariant to the details of the intensity patterns. From a related perspective, the weight vectors represent a dictionary of templates of image patterns. The final “decision layers” of the Deep Net are usually harder to interpret but it is plausible that they make decisions based on the templates represented by the lower layers. This “dictionary of templates” interpretation of Deep Nets suggests they are very efficient to learn and represent an enormous variety of image patterns, and interpolate between them, but cannot extrapolate much beyond the patterns they have seen in their training dataset. Other studies suggest that Deep Nets are less effective at modeling visual properties which are specified purely by geometry, particularly if the input consists of binary valued patterns corresponding to the presence or absence of boundary edges. It is an open issue whether Deep Nets can learn features that “factorize” different visual properties which, as we will argue later in Section 7, will ultimately be necessary for dealing with the full complexity of real images.



**Fig. 3** Figure taken from Wang et al. (2015). The visual concepts obtained by population encoding are visually tight and we can identify the parent object class pretty easily by just looking at the mid-level concepts.

## 5 Deep Nets and Biological Vision

Deep Nets have a lot to offer for studying biological vision systems and, in particular, disciplines like cognitive science, neuroscience and psychology which aim at understanding the mind and the brain. They can help develop and test computational theories by exploiting the availability of big data while raising the possibility of understanding the brain by relating the artificial neurons in Deep Nets to real neurons in the brain. But they also have significant limitations for both modeling real neural circuits and human cognitive abilities.

### 5.1 Exploiting Big Data

The use of Deep Nets, and other machine learning techniques, can help develop theories of mind and brain which exploit big data. This can be done in roughly three ways. Firstly, Deep Nets can help develop theories that deal with the enormous complexity of real world images. Secondly, they can be used to partially learn the knowledge about the visual world that humans and other animals obtain through development and experience. Thirdly, they enable theories to be tested on complex stimuli and compared to alternative theories. We will now address these issues in turn.

Historically, studies of biological visual systems have largely relied on simple synthesized stimuli. These studies have led to many important findings and were historically necessary because the complexity of natural image stimuli means that it is extremely hard to perform controlled scientific experiments by systematically varying the experimental parameters. This also follows the well established scientific strategy of divide and conquer which aims at understanding by breaking down complex phenomena into more easily understandable chunks. But studying vision on simplified stimuli has limitations which Deep Nets and big data can help address. As researchers in computer vision discovered

in the 1980s, findings on simplified synthetic stimuli, though sometimes providing motivations and good starting points, typically required enormous modifications before they could be extended to realistic stimuli if they could be extended at all. Computer vision researchers had to leave their comfort zone of synthetic stimuli and address the fundamental challenge of vision: namely how visual systems deal with the complexity and ambiguity of real world images and achieve the miracle of converting the light rays that enter the eye, or a camera, into an interpretation of the three-dimensional physical world. Driven by the need to address these issues, computer vision researchers developed a large set of mathematical and computational techniques and increasingly realized the importance of learning theories from data using tools like Deep Nets, which required large annotated datasets. The same techniques can be directly applied to studying biological vision by predicting experimental responses to visual stimuli, e.g., human performance in behavioral experiments, the responses of neurons, or fMRI activity.

Big data, and learning methods for mining the data, are particularly important for vision because, as leading vision scientists like Gregory and Marr have argued, visual systems require knowledge of the world in the form of natural and ecological constraints. In Gregory's words "perception is not just a passive acceptance of stimuli, but an active process involving memory and other internal processes". In other words, the visual systems of humans, and other animals, exploit a large amount of knowledge which has been acquired through development and experience. Big data methods, like Deep Nets, gives a surrogate way for vision scientists to partially learn this knowledge by studying properties of real world images.

Finally, the use of big datasets are also very important for testing visual theories because they enabled detailed comparisons with alternative theories. They make it easy to reject "toy theories" that exploit the

biases inherent in small datasets and simplified stimuli. In summary, the use of Deep Nets and big data enable biological vision researchers to develop and test theories that can work in realistic visual domains and address the fundamental challenge of vision.

## 5.2 Real Neurons and Neural Circuits

From the neuroscience perspective, Deep Nets have been used to predict brain activity, such as fMRI and other non-invasive measurements, and there are a growing number of examples (Cichy et al., 2016; Wen et al., 2017). They have also been applied to predicting neural responses as measured by electrophysiology and, in particular, for predicting the response of neurons in the ventral stream (Yamins et al., 2014). These are examples where Deep Nets’ ability to learn from data and to deal with the complexity of real stimuli really pays off. But in terms of understanding neuroscience, this is best thought of as a starting point. The ventral stream of primates is very complex and there is evidence that it estimates the three-dimensional structure of objects and parts (Yamane et al., 2008), and relates to the classic theory of object recognition by component (Biederman, 1987) which differs in many respects from standard Deep Nets. More generally, the primate visual systems must perform all the visual tasks listed in Section 3, namely edge detection, binocular stereo, semantic segmentation, object classification, scene classification, and 3D-depth estimation. The vision community has developed a range of different Deep Nets for these tasks so it is extremely unlikely, for example, for a Deep Net trained for object classification on ImageNet to be able to account for the richness of primate visual systems.

It should also be emphasized that while Deep Nets perform computations bottom-up in a feedforward manner there is considerable evidence of top-down processing in the brain (Lee and Mumford, 2003), particularly driven by top-down attention (Gregoriou et al., 2014). Researchers have also identified cortical circuits (McManus et al., 2011) which implement spatial interactions (though possibly in a bottom-up and top-down manner). These types of phenomena require other families of mathematical models, perhaps the compositional models described in Section 7.

But, more fundamentally, it must be acknowledged that there are big differences between the artificial neurons used in Deep Nets and real neurons in the brain. Artificial models of neurons are, at best, great simplifications of realistic neurons as shown by studies of real neurons *in vitro* (Poirazi and Mel, 2001). Neuroscientists have found that there are over one hundred dif-

ferent types of neurons, and there are enormous morphological differences which may be exploited to enable computation (Seung, 2012). There is also lack of detailed understanding of neural circuits. For example, the wiring diagram of *C-elegans* has been known for over thirty years but there is still only limited understanding of how it functions as a neural circuit (as stated by O. Hobert the wiring diagram “is like a road map that tells you where cars can drive, but does not tell you when or where cars are actually driving”). Understanding neural circuits will also require understanding their dynamics and how this can change based on a host of possible mechanisms such as rapidly changing synapses (Von Der Malsburg, 1994). Understanding real neurons and real neural circuits is a fascinating scientific challenge and exciting engineering advances (Boyden et al., 2005) and the availability of huge datasets and the tools to analyze them means that progress will surely be made. But these are highly challenging scientific tasks. In summary, the jump between real neural circuits and the artificial circuits in Deep Nets remains huge and it is likely that real neural circuits will be ultimately found to be much more complicated.

## 5.3 Cognitive Abilities: Deep Nets and Scientific Understanding

It is clear that Deep Nets, and other machine learning techniques, are very helpful for vision scientists but are doubtful that they are sufficient to capture the complexity of biological visual systems. The human visual system performs much better than Deep Nets, or other AI visual systems, on almost all visual tasks. The few exceptions are on situations for which evolution and experience put humans at a disadvantage. For example, AI systems can outperform humans by recognizing hundreds of millions of faces provided they are seen from front-on under reasonable lighting conditions and with limited occlusion, but until recently most humans never saw more than a few thousand people in their whole lifetime. It is also possible that AI systems could perform better than the average radiologists when reading computer tomography (CT) images, but even the most expert radiologists have only seen a fairly small number of CT scans (and AI systems can directly access the three-dimensional data in CT scans, while radiologists can only view two-dimensional slices). In each of these cases, humans are at a disadvantage because they do not have access to, and hence cannot exploit, the enormous amounts of annotated big data which enable Deep Nets to do so well on these tasks. But true examples of Deep Nets outperforming humans are very rare (and often due to Deep Nets overfitting the datasets on which

the studies are performed). Moreover, humans can perform a large variety of visual tasks while current AI systems are usually specialized on single tasks.

Moreover, studies of cognitive science show that human visual systems can work at levels of abstraction which current Deep Nets cannot match. This can be illustrated by human ability at visual analogies some of which depend only on visual similarity but others depend on the notion of parts and subparts, while others include the idea of function. As we will argue in Section 7 this reflects limitations of current machine learning methods and the suggestion that current techniques, like Deep Nets, will reach a wall. From another perspective, it can also be argued that the goal of vision science is to discover underlying principles. From this perspective, a model that explains phenomena in terms of an uninterpretable Deep Net would not be very satisfying. This is a debatable issue on which reasonable people can disagree. But we suspect that progress in AI will also require interpretable models partly for the pragmatic engineering principle, that this is necessary for debugging and for performance and safety guarantees.

In summary, Deep Nets, and other techniques which exploit big data, are a tool that mind and brain researchers should know how to use and not misuse. But it is equally clear that current Deep Nets fail to capture some of the most interesting phenomena such as human’s ability to perform abstractions and perform analogical reasoning (although Deep Nets might be useful as building blocks to construct such a theory). Nevertheless a closer relationship between biological and artificial models of vision would be beneficial to both disciplines. Researchers in AI have developed a large set of technical tools, like Deep Nets, which can allow their models to be applied to the complexity of natural images and tested under rigorous realistic conditions. Vision scientists can challenge computer vision researchers to develop theories which can perform as well as, or better than humans, in challenging situations while using orders of magnitude less power than current computers.

## 6 Some Challenges

This section describes some of the current challenges of Deep Nets and the attempts to address them. Some of these challenges are gradually being overcome while others, such the sensitivity to non-local attacks, may require more fundamental changes as we will discuss in Section 7.

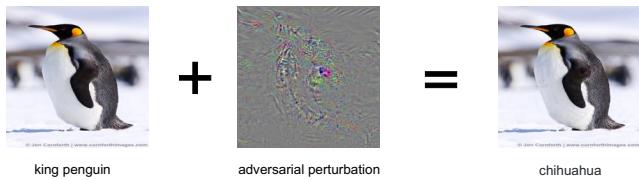
### 6.1 Relaxing the Need for Full Supervision

A disadvantage of Deep Nets is that they typically need a very large amount of annotated training data, which restricts their use to situations where big data is available. But this is not always the case. In particular, “transfer learning” shows that the features of Deep Nets learned on annotated datasets for certain visual tasks can sometimes be transferred to novel datasets and related tasks, thereby enabling learning with much less data and sometimes with less supervision. For example, as mentioned earlier, Deep Nets were first successful for object classification on ImageNet but failed on object detection on the smaller PASCAL dataset. This was presumably because PASCAL was not big enough to train a Deep Net but ImageNet was (ImageNet is almost two orders of magnitude larger than PASCAL). But researchers quickly realized that it was possible to train a Deep Net for object detection and semantic segmentation on PASCAL by initializing the weights of the Deep Net by the weights of a Deep Net trained on ImageNet (Girshick et al., 2014; Long et al., 2015; Chen et al., 2018). This also introduced a mechanism for generating proposals, see Figure 1 (bottom right).

This ability to transfer Deep Net knowledge learned on another domain relates intuitively to the way children learn. A child initially learns rather slowly compared to other young animals but at critical periods the child’s learning accelerates very rapidly (Smith and Gasser, 2005). From the “dictionary of templates” perspective, this could happen because after a child has learned to recognize enough objects he/she may have enough building blocks (i.e. deep network filters) to be able to represent new objects in terms of a dictionary of existing templates. If so, only a few examples of the new object may be needed in order to do few-shot learning.

Few-shot learning of novel object categories has been shown for Deep Nets provided they have first been trained on a large set of object categories (Mao et al., 2015; Vinyals et al., 2016; Qiao et al., 2018). Another strategy is to train a Deep Net to learn similarity (technically a *Siamese network*) on the set of object categories, hence obtaining a similarity measure for the new objects. For example, Lin et al. (2017) trained a Siamese network to learn similarity for objects in ShapeNet (Chang et al., 2015) and then this similarity measure was used to cluster objects in the Tufa dataset (Salakhutdinov et al., 2012). Other few-shot learning tasks can also be done by using features from Deep Nets trained for some other tasks as ways to model the visual patterns of objects.

More recently, there has been work on unsupervised learning which shows that optical flow and structure from motion can be learned without requiring detailed



**Fig. 4** Figure taken from Xie et al. (2018). A deep network can correctly classify the left image as *king penguin*. The middle image is the adversarial noise magnified by 10 and shifted by 128, and on the right is the adversarial example misclassified as *chihuahua*.

supervision but only an energy function model (Ren et al., 2017; Zhou et al., 2017). Like many neural nets in the third wave some of the basic ideas can be found in obscure papers from the second wave (Smirnakis and Yuille, 1995). In some cases, this can even be bootstrapped to learning depth from single images. Other forms of unsupervised learning show that Deep Net features can be learned by distinguishing between scrambled and unscrambled images (Doersch et al., 2015), or by tracking an object over time (Wang and Gupta, 2015).

Other studies show that Deep Nets can exploit large numbers of unsupervised, or weakly supervised, data provided they have sufficient annotated data to start with. For example, to train object detection using images where only the names of the objects in the image are known but their locations and sizes are unknown. This is known as weakly supervised learning and it can be treated as missing/hidden data problem which can be addressed by methods such as Multiple Instance Learning (MIL) or Expectation-Maximization (EM). Performance of these types of methods is often improved by using a small amount of fully supervised training data which helps the EM or MIL algorithms converge to good solutions, e.g., see Papandreou et al. (2015).

## 6.2 Defending Against Adversarial Examples

Another limitation of Deep Nets comes from studies showing they can be successfully attacked by imperceptible modifications of the images which nevertheless cause the Deep Nets to make major mistakes for object classification (Szegedy et al., 2014), object detection, and semantic segmentation (Xie et al., 2017) (see Figure 4 and Figure 5). This problem partly arises because the datasets are finite and contain only an infinitesimal fraction of all possible images. Hence there are infinitely many images arbitrarily close to the training images and so there is a reasonable chance that the Deep Net will misclassify some of them. Researchers have shown that they can find such images either by



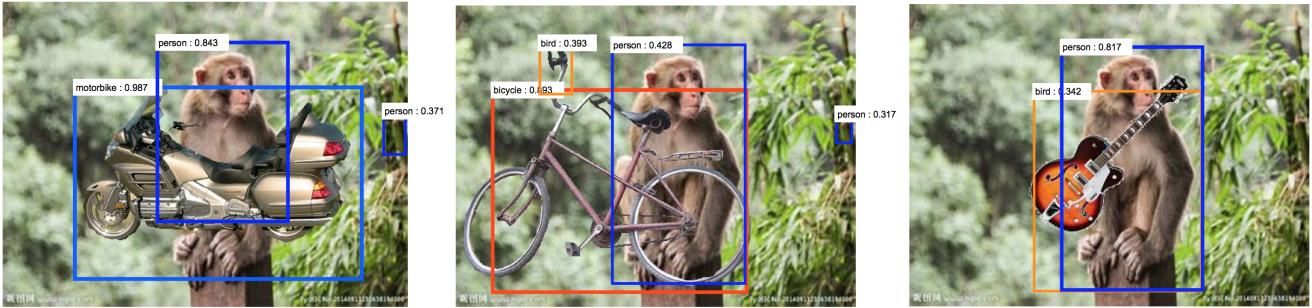
**Fig. 5** Figure taken from Xie et al. (2017). The top row is the input (adversarial perturbation already added) to the segmentation network, and the bottom row is the output. The red, blue and black regions are predicted as *airplane*, *bus* and *background*, respectively.

*white box* attacks, where the details of the Deep Net are known, or by *black box* attacks, when they are not. But there are now strategies which defend against these attacks. One strategy is to treat these “attack images” as extra training data, known as “adversarial training” (Goodfellow et al., 2015; Madry et al., 2017). A second recent alternative (Xie et al., 2018) is to introduce small random perturbations into the images, exploiting the assumption that the “attack images” are very unstable so small random perturbation will defend against them (admittedly Athalye et al. (2018) has successfully circumvented this defense). It should be acknowledged that adversarial attacks can be mounted against any vision algorithm and it would be much easier to successfully attack most other vision algorithms.

## 6.3 Addressing Over-Sensitivity to Context

A more serious challenge to Deep Nets is their over-sensitivity to context. Figure 6 shows the effect of photoshopping a guitar into a picture of a monkey in the jungle. This causes the Deep Net to misidentify the monkey as a human and also misinterpret the guitar as a bird, presumably because monkeys are less likely than humans to carry a guitar and birds are more likely than guitars to be in a jungle near a monkey (Wang et al., 2018). Recent work gives many examples of the over-sensitivity of Deep Nets to context, such as putting an elephant in a room (Rosenfeld et al., 2018).

This over-sensitivity to context can also be traced back to the limited size of datasets. For any object only a limited number of contexts will occur in the dataset and so the Deep Net will be biased towards them. For



**Fig. 6** Figure taken from Wang et al. (2018). Adding occluders cause deep network to fail. Left: The occluding motorbike turns a monkey into a human. Center: The occluding bicycle turns a monkey into a human and the jungle turns the bicycle handle into a bird. Right: The occluding guitar turns the monkey into a human and the jungle turns the guitar into a bird.



**Fig. 7** Hazardous factors for stereo vision, as identified in Zendel et al. (2015). These challenging scenarios do not systematically appear in real world, so relying on synthetic data is a promising alternative.

example, in early image captioning datasets it was observed that giraffes only occurred with trees and so the generated captions failed to mention giraffes in images without trees even if they were the most dominant object.

Observe that the limited size of datasets is a common theme when we consider the current limitations of Deep Nets. Recall that we already mentioned how synthetic data could be used, see Figure 2, to show that Deep Nets trained on ImageNet could not recognize objects from some viewpoints. An advantage of synthetic data is that it enables us to generate, in principle, an infinite amount of images and hence to systematically explore the effect of varying factors like viewpoint and material properties, e.g., see Qiu and Yuille (2016); Alcorn et al. (2018). Similarly synthetic data can be used to systematically vary hazardous factors for stereo vision (those factors like specularity which are known to cause stereo algorithms to fail; see Figure 7) enabling researchers to characterize the sensitivity of stereo algorithms to these factors (Zhang et al., 2018). Hence synthetic datasets offer the possibility of generating as much data as is required to systematically study the sensitivity of Deep Nets to the nuisance factors, like viewpoint and radiosity, which arrive in reality (provided the synthetic datasets are realistic enough to accurately represent real world images).

The difficulty of capturing the enormous varieties of context, as well as the need to explore the large range of nuisance factors, is highly problematic for data driven methods like Deep Nets. It seems that ensuring that the networks can deal with all these issues will require datasets that are arbitrarily big, which raises enormous challenges for both training and testing datasets. We will discuss these issues next.

## 7 The Combinatorial Explosion: When Big Datasets Are Not Enough

This section argues that vision researchers face a combinatorial explosion as they grapple with the complexity of real world data in order to develop algorithms that will work robustly on complex visual tasks in the real world. In such situations big datasets will not be big enough and novel methods will be required for developing algorithms and for testing them.

### 7.1 The Combinatorial Explosion

Deep Nets are trained and evaluated on large datasets which are intended to be representative of the real world. But, as discussed earlier, Deep Nets can fail to generalize to images outside the datasets they were trained on, can make mistakes on rare events that occur rarely



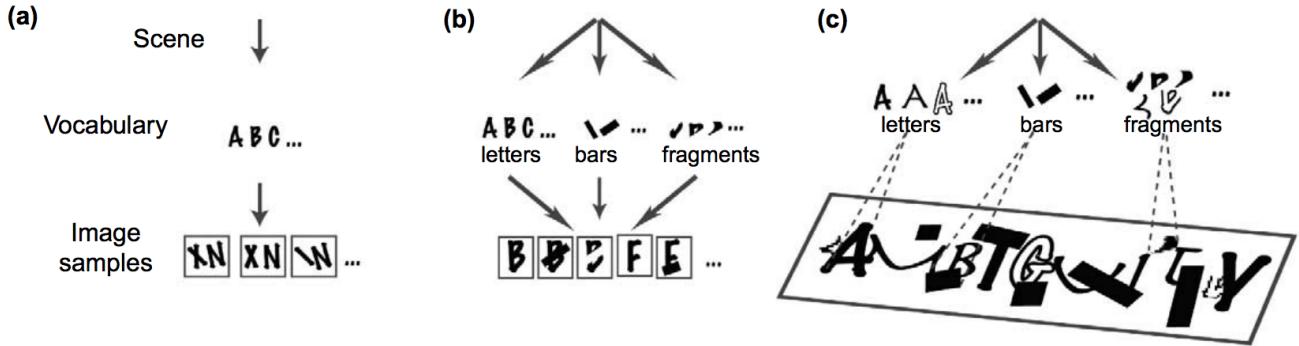
**Fig. 8** An illustration of combinatorial explosion. We consider the (already simplified) rendering process of one object. It involves choosing the camera pose, lighting condition, object texture, etc: a total of (merely) 13 parameters. If we allow 1,000 different values for each parameter, then we obtain a total of  $10^{39}$  different images. This is way beyond the size of any dataset, as well as the number of images humans see per year.

within the datasets (but which may have disastrous consequences, such as running over a baby or failure to detect a cancerous tumor), and are also sensitive to adversarial attacks and changes in context. None of these problems are necessarily deal-breakers for the success of Deep Nets and they can certainly be overcome for certain visual domains and tasks. But we argue that these are early warning signs of a problem that will arise as vision researchers attempt to use Deep Nets to address increasingly complex visual tasks in unconstrained domains. Namely, that in order to deal with the combinatorial complexity of real world images the datasets would have to become exponentially large, which is clearly impractical.

To understand this combinatorial complexity consider the following thought experiment. Imagine constructing a visual scene by selecting objects from an object dictionary and placing them in different configurations. This can clearly be done in an exponential number of ways. We can obtain similar complexity even for images of a single object since it can be partially occluded in an exponential number of ways. We can also change the context of an object in an infinite number of ways. Although humans are good at adapting to changes in visual context, Deep Nets are much more sensitive, as illustrated in Figure 6. We note that this combinatorial explosion may not happen for some visual tasks and Deep Nets are likely to be extremely successful for medical image application because there is comparatively little variability in context (e.g., the Pancreas is always very close to the Duodenum). But for many real world applications, particularly those involving humans interacting with the world in video sequences, it seems that the complexity of the real world cannot be captured without having an exponentially large dataset.

This causes big challenges for current methods of training and testing visual algorithms. These methods were developed by machine learning researchers to ensure that algorithms are capturing the underlying structure of the data instead of merely memorizing the training data. They assume that the training and testing data are randomly drawn samples from some unknown probability distributions. But critically, the datasets need to be large enough to be representative of the underlying distribution of the data. Interestingly, to the best of our knowledge, researchers on the foundations of machine learning have never directly addressed this issue. Instead they have concentrated on theoretical results, called Probably Approximately Correct (PAC) theorems, which give bounds on the probability that a machine learning algorithm has learned the structure of the underlying data, whose key insight is that the amount of training data must be much larger than the set of hypotheses that the learning algorithm can consider before seeing the data (Valiant, 1984; Vapnik, 1998; Poggio and Smale, 2003). But, in any case, the standard paradigm of training and testing models on a finite number of randomly drawn samples becomes impractical if the set of images is combinatorially large. This forces us to address two new problems: (I) How can we train algorithms on finite sized datasets so that they can perform well on the truly enormous datasets required to capture the combinatorial complexity of the real world? (II) How can we efficiently test these algorithms to ensure that they work in these enormous datasets if we can only test them on a finite subset?

It helps to consider these issues from the perspective of computer graphics. It is straightforward (see Figure 8) to specify a computer program with 13 parameters that can render images of a single object from different viewpoints, under different illuminations, and



**Fig. 9** Figure taken from Yuille and Kersten (2006). From (a) to (b) to (c), an increasing level of variability and occlusion is used, yet humans can still do inference and correctly interpret the image.

in a limited number of background scenes. If we allow 1,000 different values for each parameter we obtain a total of  $10^{39}$  different images,  $10^{30}$  orders of magnitude larger than any existing dataset. The program can be extended to include multiple objects in an enormous range of visual scenes and, in principle, we can specify a model with a finite, but very large, number of parameters that can generate a combinatorially large number of real images which can approximate the real world. But while this gives a way to potentially generate all real world images it does not solve the issue of how to train and test models on these datasets.

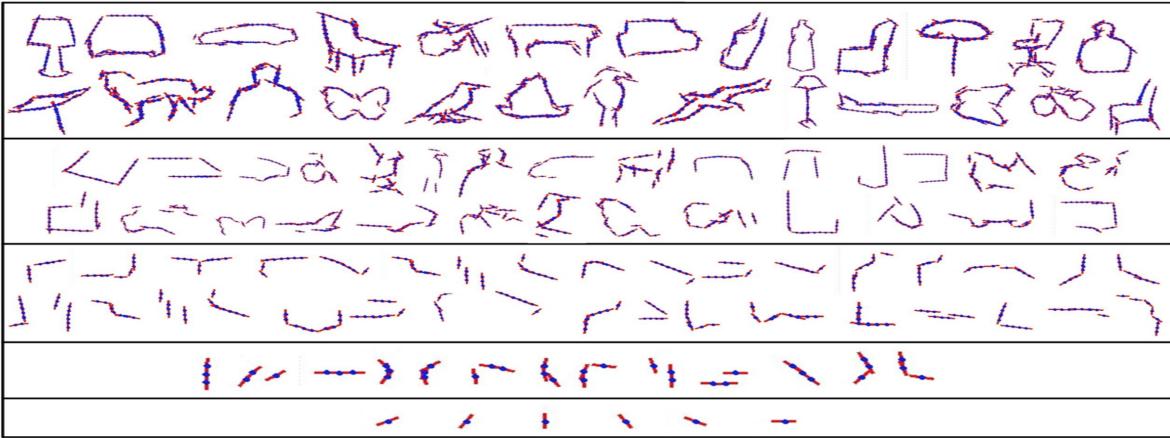
## 7.2 Models for Overcoming Combinatorial Complexity

It seems highly unlikely that methods like Deep Nets, in their current forms, can deal with the combinatorial explosion. The datasets may never be large enough to either train or test them. Here we sketch the types of ideas we think will be relevant. We can get some guidance from the human visual system which faces and overcomes these challenges. Humans see roughly  $10^9$  images every year (assuming 30 images per second) which is big, but not combinatorial. But humans, above a critical age, can learn from small numbers of examples, perceive three-dimensional structure, deal with abstraction, can exploit context when it is helpful but ignore it when it is not. Recent experiments (Ullman et al., 2016) suggest that humans can interpret images unambiguously provided they are above a critical size (which depends on the image content) and additional context is unnecessary.

Compositionality will probably be one part of the solution. This is a general principle which can be described poetically as “an embodiment of faith that the world is knowable, that one can tease things apart, comprehend them, and mentally recompose them at will”. The key assumption is that structures are composed

hierarchically from more elementary substructures following a set of grammatical rules. This suggests that the substructures and the grammars can be learned from finite amounts of data but will generalize to combinatorial situations. Unlike Deep Nets, compositional models require structured representations which make explicit their structures and substructures which enables them to do multiple tasks (e.g., detecting objects, object parts, and object boundaries) with the same underlying representation (Chen et al., 2007) (it is argued that Deep Nets are compositional, but this is in a very different sense). Compositional models offer the ability to extrapolate beyond data they have seen, to reason about the system, intervene, do diagnostics, and to answer many different questions based on the same underlying knowledge structure (Pearl, 2009). To quote Stuart Geman “the world is compositional or God exists”, since otherwise it would seem necessary for God to hardwire human intelligence (Geman, 2007).

Compositionality relates closely to pattern theory and analysis by synthesis (Grenander, 1993; Mumford, 1994; Tu et al., 2003; Zhu and Mumford, 2006; Mumford and Desolneux, 2010). It can be illustrated by a toy-world example, shown in Figure 9, where images are created in terms of basic vocabularies of elementary components. The three panels show microworlds of increasing complexity from left to right. For each microworld there is a grammar which specifies the possible images as constructed by compositions of the elementary components. In the left panel the elementary components are letters which do not overlap, and so interpreting the image is easy. The center and right panels are generated by more complicated grammars – letters of different fonts, bars, and fragments which can heavily occlude each other. Interpreting these images is much harder and seems to require the notion that letters are composed of elementary parts, that they can occur in a variety of fonts, and the notion of “explaining away”



**Fig. 10** Figure taken from Zhu et al. (2010). Mean shapes from Recursive Compositional Models at different levels. This hierarchy was learned in an unsupervised manner.

(to explain that parts of a letter are missing because they have been occluded by another letter).

The third microworld in Figure 9 is an example of a combinatorially large dataset since images are constructed by selecting objects from a dictionary and placing them at random while allowing for occlusion. This microworld is essentially the same as CAPTCHAs which can be used to distinguish between humans and robots. Interestingly, work on CAPTCHAs (George et al., 2017) show that compositional models which represent objects in terms of compositions of elementary tokens and factorize geometry and appearances can perform well on these types of datasets. Their inference algorithm involves bottom-up and top-down processing (Tu et al., 2003) which enables the algorithm to “explain away” missing parts of the letters and to impose “global consistency” of the interpretation to remove ambiguities. Intuitively, part detectors make bottom-up proposals for letters which can be validated or rejected in the top-down stage. By contrast, Deep Nets performed much worse on these datasets. Presumably because, unlike compositional models, they cannot capture the underlying generative structure of the domain and extrapolate outside their training dataset. Since the microworld is combinatorially large, it will not be possible to train Deep Nets on enough data to guarantee good performance on the entire dataset. Other theoretical studies, e.g., Yuille and Mottaghi (2016), suggest that compositional models are well suited for dealing with complexity by sharing parts and using hierarchical abstraction.

Other non-visual examples illustrate the same points. A recent example is when researchers (Santoro et al., 2018) tried to train standard Deep Nets to do IQ tests. The task requires finding composition of meaningful rules/patterns (distractors may be present) within 8 given images in a  $3 \times 3$  grid, and the goal is to fill

in the last missing image. Not surprisingly, Deep Nets do not generalize well. For natural language applications, Neural Module Networks (Andreas et al., 2016) are more promising than static, fixed-structure Deep Nets, in that the dynamic architectural layout may be flexible enough to capture some meaningful compositions. In fact, we recently verified that the individual modules indeed perform their intended functionalities (e.g. AND, OR, Filter(red) etc) after joint training (Liu et al., 2019).

Compositional models have many desirable theoretical properties, such as being *interpretable*, and the ability to be *generative* so they can be sampled from. This means that, in principle, they know everything about the object (or whatever entity is being modeled) which makes them easier to diagnose, and hence harder to fool, than black box methods like Deep Nets. But learning compositional models is hard because it requires learning the building blocks and the grammars (and even the nature of the grammars is debatable). There has, however, been some limited success in learning hierarchical dictionaries starting from basic elementary tokens like edges (Zhu et al., 2010): see Figure 10.

A current limitation of compositional models is that in order to perform analysis by synthesis they need to have generative models of objects and scene structures. Putting distributions on images is challenging with a few exceptions like faces, letters, and regular textures (Tu et al., 2003). But there is promising progress from two directions. Firstly, computer graphics models are becoming increasingly realistic and visual appearance can be roughly factored into geometry, texture, and illumination. Recall that the  $10^{39}$  images (Figure 8) were generated from only 13 parameters. Secondly, Deep Nets have also been applied to generating images using Generative Adversarial Networks (GANs).

From the perspective of analysis by synthesis, the results of GANs are disappointing though recent work on conditional GANs shows promise.

More fundamentally, dealing with the combinatorial explosion requires learning causal models of the 3D world and how these generate images. Studies of human infants suggest that they learn by making causal models that predict the structure of their environment including naive physics. This causal understanding enables learning from limited amounts of data and performing true generalization to novel situations. This is analogous to contrasting Newton’s Laws, which gave causal understanding with a minimal amount of free parameters, with the Ptolemaic model of the solar system gave very accurate predicts but required a large amount of data to determine its details (i.e. the epicycles).

### 7.3 Testing Models When Data Is Combinatorial

How can we test vision algorithms to deal with the complexity of the real world if we can only test them on finite amounts of data? If we have well structured models, e.g., compositional models as described above, then we can exploit the structure of the models to determine their failure modes. This, of course, is similar to how complex engineering (e.g., airplanes) or software structures are tested by systematically identifying their weak points. This is more reminiscent of game theory rather than decision theory (which focuses on the average loss and which underlies machine learning theory) because it suggests paying attention to the worst cases instead of the average cases. This makes sense if the goal is to develop visual algorithms for self-driving cars, or diagnosing cancer in medical images, where failures of the algorithms can have major consequences.

This can be done already if the failure modes of the visual tasks can be identified and are low-dimensional. For example, as mentioned earlier in Section 6.3, researchers have isolated the hazardous factors which cause stereo algorithms to fail which include specularities and texture-less regions. In such cases it is possible to exploit computer graphics to systematically vary these hazardous factors to determine which algorithms are resistant to them (Zhang et al., 2018). In short, we can stress-test these algorithms along these specific dimensions.

But for most visual tasks it is very hard to identify a small number of hazard factors which can be isolated and tested further. Instead, we should generalize the notion of adversarial attacks to include non-local structure. A simple possibility is to allow other more complex operations which cause reasonable changes to the image or scene, e.g., by occlusion, or changing the

physical properties of the objects being viewed (Zeng et al., 2017), but without significantly impacting human perception.

## 8 Conclusion

This opinion piece has been motivated by discussions about Deep Nets with researchers in many different disciplines. We have tried to strike a balance which acknowledges the immense success of Deep Nets but which does not get carried away by the popular excitement surrounding them. We have often used work from our own group to illustrate some of our main points and apologize to other authors whose work we would have cited in a more scholarly review of the field. Several of our concerns parallel those mentioned in recent critiques of Deep Nets (Darwiche, 2018; Marcus, 2018).

A few years ago Aude Oliva and the first author co-organized a NSF-sponsored workshop on the Frontiers of Computer Vision (MIT CSAIL, August 21-24 2011). The meeting encouraged frank exchanges of opinion and, in particular, there was enormous disagreement about the potential of Deep Nets for computer vision. But a few years later, as Yann LeCun predicted, everybody is using Deep Nets. Their successes have been extraordinary and have helped vision become much more widely known, dramatically increased the interaction between academia and industry, lead to application of vision techniques to a large range of disciplines, and have many other important consequences. But despite their successes there remain enormous challenges which must be overcome before we reach the goal of general purpose artificial intelligence and understanding of biological vision systems. In particular dealing with the combinatorial explosion as researchers address increasingly complex visual tasks in real world conditions. While Deep Nets, and other big data methods, will surely be part of the solution we believe that we will also need complementary approaches which can build on their successes and insights.

**Acknowledgements** This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and ONR N00014-15-1-2356. We thank Kyle Rawlins and Tal Linzen for providing feedback.

## References

- Alcorn MA, Li Q, Gong Z, Wang C, Mai L, Ku W, Nguyen A (2018) Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. CoRR abs/1811.11553

- Andreas J, Rohrbach M, Darrell T, Klein D (2016) Neural module networks. In: CVPR, IEEE Computer Society, pp 39–48
- Athalye A, Carlini N, Wagner DA (2018) Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: ICML, JMLR.org, JMLR Workshop and Conference Proceedings, vol 80, pp 274–283
- Biederman I (1987) Recognition-by-components: a theory of human image understanding. *Psychological review* 94(2):115
- Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K (2005) Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience* 8(9):1263
- Chang AX, Funkhouser TA, Guibas LJ, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, Xiao J, Yi L, Yu F (2015) Shapenet: An information-rich 3d model repository. CoRR abs/1512.03012
- Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell* 40(4):834–848
- Chen Y, Zhu L, Lin C, Yuille AL, Zhang H (2007) Rapid inference on a novel AND/OR graph for object detection, segmentation and parsing. In: NIPS, Curran Associates, Inc., pp 289–296
- Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports* 6:27755
- Darwiche A (2018) Human-level intelligence or animal-like abilities? *Commun ACM* 61(10):56–67
- Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) Imagenet: A large-scale hierarchical image database. In: CVPR, IEEE Computer Society, pp 248–255
- Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: ICCV, IEEE Computer Society, pp 1422–1430
- Everingham M, Gool LJV, Williams CKI, Winn JM, Zisserman A (2010) The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision* 88(2):303–338
- Felzenszwalb PF, Girshick RB, McAllester DA, Ramanan D (2010) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32(9):1627–1645
- Fukushima K, Miyake S (1982) Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and co-operation in neural nets, Springer, pp 267–285
- Geman S (2007) Compositionality in vision. In: The grammar of vision: probabilistic grammar-based models for visual scene understanding and object categorization
- George D, Lehrach W, Kansky K, Lázaro-Gredilla M, Laan C, Marthi B, Lou X, Meng Z, Liu Y, Wang H, et al. (2017) A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science* 358(6368):eaag2612
- Girshick RB, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, IEEE Computer Society, pp 580–587
- Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: International Conference on Learning Representations
- Gregoriou GG, Rossi AF, Ungerleider LG, Desimone R (2014) Lesions of prefrontal cortex reduce attentional modulation of neuronal responses and synchrony in v4. *Nature neuroscience* 17(7):1003–1011
- Grenander U (1993) General pattern theory-A mathematical study of regular structures. Clarendon Press
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, IEEE Computer Society, pp 770–778
- Hornik K, Stinchcombe MB, White H (1989) Multi-layer feedforward networks are universal approximators. *Neural Networks* 2(5):359–366
- Kokkinos I (2017) Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: CVPR, IEEE Computer Society, pp 5454–5463
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS, pp 1106–1114
- Lee TS, Mumford D (2003) Hierarchical bayesian inference in the visual cortex. *JOSA A* 20(7):1434–1448
- Lin X, Wang H, Li Z, Zhang Y, Yuille AL, Lee TS (2017) Transfer of view-manifold learning to similarity perception of novel objects. In: International Conference on Learning Representations
- Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li L, Fei-Fei L, Yuille AL, Huang J, Murphy K (2018) Progressive neural architecture search. In: ECCV (1), Springer, Lecture Notes in Computer Science, vol 11205, pp 19–35
- Liu R, Liu C, Bai Y, Yuille A (2019) Clevr-ref+: Diagnosing visual reasoning with referring expressions. CoRR abs/1901.00850
- Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: CVPR, IEEE Computer Society, pp 3431–3440

- Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. CoRR abs/1706.06083
- Mao J, Wei X, Yang Y, Wang J, Huang Z, Yuille AL (2015) Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In: ICCV, IEEE Computer Society, pp 2533–2541
- Marcus G (2018) Deep learning: A critical appraisal. CoRR abs/1801.00631
- McManus JN, Li W, Gilbert CD (2011) Adaptive shape processing in primary visual cortex. Proceedings of the National Academy of Sciences 108(24):9739–9746
- Mumford D (1994) Pattern theory: a unifying perspective. In: First European congress of mathematics, Springer, pp 187–224
- Mumford D, Desolneux A (2010) Pattern theory: the stochastic analysis of real-world signals. CRC Press
- Papandreou G, Chen L, Murphy KP, Yuille AL (2015) Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV, IEEE Computer Society, pp 1742–1750
- Pearl J (2009) Causality. Cambridge university press
- Poggio T, Smale S (2003) The mathematics of learning: Dealing with data. Notices of the AMS 50(5):537–544
- Poirazi P, Mel BW (2001) Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. Neuron 29(3):779–796
- Qiao S, Liu C, Shen W, Yuille AL (2018) Few-shot image recognition by predicting parameters from activations. In: CVPR, IEEE Computer Society, pp 7229–7238
- Qiu W, Yuille AL (2016) Unrealcv: Connecting computer vision to unreal engine. In: ECCV Workshops (3), Lecture Notes in Computer Science, vol 9915, pp 909–916
- Ren S, He K, Girshick RB, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, pp 91–99
- Ren Z, Yan J, Ni B, Liu B, Yang X, Zha H (2017) Unsupervised deep learning for optical flow estimation. In: AAAI, AAAI Press, pp 1495–1501
- Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nature neuroscience 2(11):1019
- Rosenfeld A, Zemel RS, Tsotsos JK (2018) The elephant in the room. CoRR abs/1808.03305
- Salakhutdinov R, Tenenbaum JB, Torralba A (2012) One-shot learning with a hierarchical nonparametric bayesian model. In: ICML Unsupervised and Transfer Learning, JMLR.org, JMLR Proceedings, vol 27, pp 195–206
- Santoro A, Hill F, Barrett DGT, Morcos AS, Lillicrap TP (2018) Measuring abstract reasoning in neural networks. In: ICML, JMLR.org, JMLR Workshop and Conference Proceedings, vol 80, pp 4477–4486
- Seung S (2012) Connectome: How the brain's wiring makes us who we are. HMH
- Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations
- Smirnakis SM, Yuille AL (1995) Neural implementation of bayesian vision theories by unsupervised learning. In: The Neurobiology of Computation, Springer, pp 427–432
- Smith L, Gasser M (2005) The development of embodied cognition: Six lessons from babies. Artificial life 11(1-2):13–29
- Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R (2014) Intriguing properties of neural networks. In: International Conference on Learning Representations
- Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: CVPR, IEEE Computer Society, pp 1521–1528
- Tu Z, Chen X, Yuille AL, Zhu SC (2003) Image parsing: Unifying segmentation, detection, and recognition. In: ICCV, IEEE Computer Society, pp 18–25
- Ullman S, Assif L, Fetaya E, Harari D (2016) Atoms of recognition in human and computer vision. Proceedings of the National Academy of Sciences 113(10):2744–2749
- Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142
- Vapnik V (1998) Statistical learning theory. Wiley
- Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D (2016) Matching networks for one shot learning. In: NIPS, pp 3630–3638
- Von Der Malsburg C (1994) The correlation theory of brain function. In: Models of neural networks, Springer, pp 95–119
- Wang J, Zhang Z, Premachandran V, Yuille AL (2015) Discovering internal representations from object-cnns using population encoding. CoRR abs/1511.06855
- Wang J, Zhang Z, Xie C, Zhou Y, Premachandran V, Zhu J, Xie L, Yuille A (2018) Visual concepts and compositional voting. Annals of Mathematical Sciences and Applications 2(3):4
- Wang X, Gupta A (2015) Unsupervised learning of visual representations using videos. In: ICCV, IEEE Computer Society, pp 2794–2802
- Wen H, Shi J, Zhang Y, Lu KH, Cao J, Liu Z (2017) Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral Cortex pp 1–25
- Xie C, Wang J, Zhang Z, Zhou Y, Xie L, Yuille AL (2017) Adversarial examples for semantic segmenta-

- tion and object detection. In: ICCV, IEEE Computer Society, pp 1378–1387
- Xie C, Wang J, Zhang Z, Ren Z, Yuille AL (2018) Mitigating adversarial effects through randomization. In: International Conference on Learning Representations
- Xu L, Krzyzak A, Yuille AL (1994) On radial basis function nets and kernel regression: Statistical consistency, convergence rates, and receptive field size. *Neural Networks* 7(4):609–628
- Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience* 11(11):1352–1360
- Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111(23):8619–8624
- Yosinski J, Clune J, Nguyen AM, Fuchs TJ, Lipson H (2015) Understanding neural networks through deep visualization. *CoRR* abs/1506.06579
- Yuille A, Kersten D (2006) Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences* 10(7):301–308
- Yuille AL, Mottaghi R (2016) Complexity of representation and inference in compositional models with part sharing. *Journal of Machine Learning Research* 17:11:1–11:28
- Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: ECCV (1), Springer, Lecture Notes in Computer Science, vol 8689, pp 818–833
- Zendel O, Murschitz M, Humenberger M, Herzner W (2015) CV-HAZOP: introducing test data validation for computer vision. In: ICCV, IEEE Computer Society, pp 2066–2074
- Zeng X, Liu C, Wang Y, Qiu W, Xie L, Tai Y, Tang C, Yuille AL (2017) Adversarial attacks beyond the image space. *CoRR* abs/1711.07183
- Zhang Y, Qiu W, Chen Q, Hu X, Yuille AL (2018) Unrealstereo: Controlling hazardous factors to analyze stereo vision. In: 3DV, IEEE Computer Society, pp 228–237
- Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A (2015) Object detectors emerge in deep scene cnns. In: International Conference on Learning Representations
- Zhou T, Brown M, Snavely N, Lowe DG (2017) Unsupervised learning of depth and ego-motion from video. In: CVPR, IEEE Computer Society, pp 6612–6619
- Zhu L, Chen Y, Torralba A, Freeman WT, Yuille AL (2010) Part and appearance sharing: Recursive compositional models for multi-view. In: CVPR, IEEE Computer Society, pp 1919–1926
- Zhu S, Mumford D (2006) A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision* 2(4):259–362
- Zhu Z, Xie L, Yuille AL (2017) Object recognition with and without objects. In: IJCAI, ijcai.org, pp 3609–3615