Searching for Computer Vision North Stars

Li Fei-Fei & Ranjay Krishna

Computer vision is one of the most fundamental areas of artificial intelligence research. It has contributed to the tremendous progress in the recent deep learning revolution in AI. In this essay, we provide a perspective of the recent evolution of object recognition in computer vision, a flagship research topic that led to the breakthrough data set of ImageNet and its ensuing algorithm developments. We argue that much of this progress is rooted in the pursuit of research "north stars," wherein researchers focus on critical problems of a scientific discipline that can galvanize major efforts and groundbreaking progress. Following the success of ImageNet and object recognition, we observe a number of exciting areas of research and a growing list of north star problems to tackle. This essay recounts the brief history of ImageNet, its related work, and the follow-up progress. The goal is to inspire more north star work to advance the field, and AI at large.

Artificial intelligence is a rapidly progressing field. To many of its everyday users, AI is an impressive feat of engineering derived from modern computer science. There is no question that there has been incredible engineering progress in AI, especially in recent years. Successful implementations of AI are all around us, from email spam filters and personalized retail recommendations to cars that avoid collisions in an emergency by autonomously braking. What may be less obvious is the science behind the engineering. As researchers in the field, we have a deep appreciation of both the engineering and the science and see the two approaches as deeply intertwined and complementary. Thinking of AI, at least in part, as a scientific discipline can inspire new lines of thought and inquiry that, in time, will make engineering progress more likely. As in any science, it is not always obvious what problems in AI are the most important to tackle. But once you have formulated a fundamental problem – once you have identified the next "north star" – you can start pushing the frontier of your field. That has certainly been our experience, and it is why we love Einstein's remark that "The mere formulation of a problem is often far more essential than its solution."

AI has been driven by north stars from the field's inception in 1950, when Alan Turing neatly formulated the problem of how to tell if a computer deserves to be called intelligent. (The computer, according to the now-famous Turing Test,

https://doi.org/10.1162/DAED_a_01902

would need to be able to "deceive a human into believing that it was human," as Turing put it.)[1] A few years later, as the founding fathers of AI planned the Dartmouth workshop, they set another ambitious goal, proposing to build machines that can "use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[2] Without that guiding light, we might never be in a position to tackle new problems.

Our own area within AI, computer vision, has been driven by its own series of north stars. This is the story of one – object recognition – and the progress it has made toward north stars in other AI fields.

The ability to see – vision – is central to intelligence. Some evolutionary biologists have hypothesized that it was the evolution of eyes in animals that first gave rise to the many different species we know today, including humans.[3]

Seeing is an immensely rich experience. When we open our eyes, the entire visual world is immediately available to us in all its complexity. From registering shadows and brightness, to taking in the colors of everything around us, to recognizing an appetizing banana on a kitchen counter as something good to eat, humans use our visual perception to navigate the world, to make sense of it, and to interact with it. So how do you even begin to teach a computer to see? There are many important problems to solve and choosing them is an essential part of the scientific quest for computer vision: that is, the process of identifying the north stars of the field. At the turn of the century, inspired by a large body of important work prior to ours, our collaborators and we were drawn to the problem of object recognition: a computer's ability to correctly identify what appears in a given image.

This seemed like the most promising north star for two reasons. The first was its practical applications. The early 2000s witnessed an explosive increase in the number of digital images, thanks to the extraordinary growth of the Internet and digital cameras, and all those images created a demand for tools to automatically catalog personal photo collections and to enable users to search through such image collections. Both applications would require object recognition.

But an even deeper reason was the remarkable ability of humans to perceive and interpret objects in the visual world. Research in the field of cognitive neuroscience showed that humans can detect animals within just twenty milliseconds and, within only three hundred milliseconds, can tell whether the animal is, say, a tiger or a lamb. The research in cognitive neuroscience also offered clues to how humans are able to achieve such rapid recognition: scientists had found that humans relied on cues in the object's surroundings and on certain key features of objects, features that did not change with a difference in angle or lighting conditions. Most strikingly, neuroscientists had discovered specific regions of the brain that activate when people view specific objects.[4] The existence of neural correlates for

any function is a sure sign of the function's evolutionary importance: a specific brain region would not evolve for a specific function unless that function was essential for the organism's survival or reproduction. Clearly, the ability to recognize specific objects must be critical.

These findings made clear to us that object recognition should be considered a north star in computer vision. But how do you get a computer to recognize objects? Recognizing objects requires understanding what concept a digital image represents in the visual world – what the image *means* – but a computer has no such understanding. To a computer, a digital image is nothing more than a collection of pixels, a two-dimensional array of numbers that does not really mean anything except colors and illuminations. Teaching a computer to recognize objects requires somehow getting it to connect each lifeless collection of numbers to a meaningful concept, like dog or banana.

Between the decades of the 1990s and the early 2000s, researchers in object recognition had already made tremendous progress toward this daunting goal, but progress was slow because of the enormous variety in the appearance of real-world objects. Even within a single, fairly specific category (like house, dog, or flower), objects can look quite different. For example, an AI capable of accurately recognizing an object in a photograph as a dog needs to recognize it as a dog whether it is a German shepherd, poodle, or chihuahua. And whatever the breed, the AI needs to recognize it as a dog whether it is photographed from the front or from the side, running to catch a ball or standing on all fours with a blue bandana around its neck. In short, there is a bewildering diversity of images of dogs, and past attempts at teaching computers to recognize such objects failed to cope with this diversity.

One major bottleneck of most of these past methods was their reliance on hand-designed templates to capture the essential features of an object, and the lack of exposure to a vast variety of images. Computers learn from being exposed to examples; that is the essence of machine learning. And while humans can often generalize correctly from just a few examples, computers need large numbers of examples; otherwise, they make mistakes. So AI researchers had been trapped in a dilemma. On the one hand, for a template to be helpful in teaching an AI system to recognize objects, the template needed to be based upon a large variety of images and, therefore, a very large number of images in total. On the other hand, hand-designing a template is labor-intensive work, and doing so from a very large number of images is not feasible.

The inability to scale the template approach effectively made it clear that we needed a different way to approach the object-recognition problem.

We started our search for a new approach with one key assumption: even the best algorithm would not generalize well if the data it learned from did not reflect the real world. In concrete terms, that meant that ma-

jor advances in object recognition could occur only from access to a large quantity of diverse, high-quality training data. That assumption may sound obvious because we are all awash in data and we all benefit from powerful object-recognition tools. But when we began our work in the early 2000s, the focus on data was fairly contrarian: at that time, most people in our field were paying attention to models (algorithms), not to data. Of course, in truth, the two pursuits are compatible. We believed that good data would help with the design of good models, which would lead to advances in object recognition and in AI more broadly.

That meant that we needed to create a new data set (which we called ImageNet) that achieved these three design goals: scale (a large quantity of data), diversity (a rich variety of objects), and quality (accurately labeled objects).[5] In focusing on these three goals, we had moved from a general north star – image recognition – to more specific problem formulations. But how did we tackle each?

Scale. Psychologists have posited that human-like perception requires exposure to thousands of diverse objects.[6] When young children learn naturally, their lives have already been exposed to enormous numbers of images every day. For example, by the time a typical child is six years old, she has seen approximately three thousand distinct objects, according to one estimate; from those examples, the child would have learned enough distinctive features to help distinguish among thirty thousand more categories. That is how large a scale we had in mind. Yet the most popular object-recognition data set when we began included only twenty objects, the result of the very process we described earlier as too cumbersome to scale up. Knowing that we needed far more objects, we collected fifteen million images from the Internet.

But images alone would not be enough to provide useful training data to a computer: we would also need meaningful categories for labeling the objects in these images. After all, how can a computer know that a picture of a dog is a German shepherd (or even a dog) unless the picture has been labeled with one of these categories? Furthermore, most of the machine learning algorithms require a training phase during which the algorithms must learn from labeled examples (that is, training examples) and be measured by their performances of a separate set of labeled examples (that is, testing samples). So we turned to an English-language vocabulary data set, called WordNet, developed by cognitive psychologist George Miller in 1990.[7] WordNet organizes words into hierarchically nested categories (such as dog, mammal, and animal); using WordNet, we chose thousands of object categories that would encompass all the images we had found. In fact, we named our data set ImageNet by analogy with WordNet.

Diversity. The images we collected from the Internet represented the diversity in real-world objects, covering many categories. For example, there were more than eight hundred different kinds of birds alone, with several examples of each. In total, we used 21,841 categories to organize the fifteen million images in our

data set. The challenges in capturing real-world diversity within each category is that simple Internet search results are biased toward certain kinds of images: for example, Google's top search results for "German shepherd" or "poodle" consist of cleanly centered images of each breed. To avoid this kind of bias, we had to expand the query to include a description: to search also, for example, for "German shepherd in the kitchen." Similarly, to get a broader, more representative distribution of the variety of dog images, we used translations into some other languages as well as hypernyms and hyponyms: not just "husky" but also "Alaskan husky" and "heavy-coated Arctic sled dog."

Quality. We cared a lot about the quality of the images and the quality of the annotations. To create a gold-standard data set that would replicate the acuity of human vision, we used only high-resolution images. And to create accurate labels for the objects in the data set, we hired people. At first, we brought in Princeton undergraduate students to label the images and verify these labels, but it quickly became apparent that using such a small group would take far too long. Through a fortunate coincidence, Amazon had just released its crowdsourcing platform, Mechanical Turk, which enabled us to quickly hire approximately fifty thousand workers from 167 countries to label and verify the objects in our set between 2007 and 2009.[8]

The ImageNet team believed it was important to democratize research in object recognition and to build a community around ImageNet. So we open-sourced ImageNet: we made it free and open to any interested researcher. We also established an annual competition to inspire researchers from all around the world. The ImageNet Large-Scale Visual Recognition Challenge (often simply called the ImageNet Challenge), which ran concurrently from 2010 until 2017 with the international computer vision research conferences International Conference on Computer Vision and European Conference on Computer Vision, created a common benchmark for measuring progress.

We set up the ImageNet Challenge similar to the design of other machine learning competitions: All participants would get the same training data, which is just a subset of the larger ImageNet data set. After using this training data to train their object-recognition algorithm, the participants would unleash their algorithm on unlabeled images that the algorithm had never encountered to see how accurately the algorithm would recognize these new images. These test data, too, came from ImageNet.

We had high aspirations for the ImageNet data set and for the ImageNet Challenge, yet the outcomes exceeded them. The biggest turning point came in 2012, when one team applied a convolutional neural network to object recognition for the first time.[9] (A convolutional neural network is an algorithm inspired by the way the human brain works.) That team's winning entry, later known as AlexNet

after one of its creators, trounced its competition, recognizing images with an accuracy that was a whopping 41 percent higher than that of the second-place finisher. Although neural networks as an approach to machine learning had been around for decades, it had not been widely used until that year's ImageNet Challenge.

This was a watershed moment for the AI community. The impressive performance of AlexNet on the ImageNet data set inspired other researchers – and not just participants in the ImageNet Challenge – to shift to deep learning approaches. We started seeing large companies like Google and Facebook deploying technology based on neural networks, and within a year, almost every AI paper was about neural networks.

With so many people working on neural networks, the technology advanced rapidly. Researchers found that the deeper the model, the better it performed at object recognition. And as deeper models required more processing power, researchers ran into other problems, such as computational bottlenecks, which required further design work to overcome. The ImageNet Challenge created a kind of domino effect of innovations, with each advance leading to more.[10]

Beyond the tremendous progress in computer vision through more and more powerful deep learning algorithms, researchers began using deep learning to automate and systematize the design of model architecture itself, instead of hand-designing each neural network's architecture. The process of hand-designing architectures, like the previous process of hand-designing features in templates, is speculative: the search space of possible architectures is exponentially vast, so manual architectural changes are unlikely to thoroughly explore this space quickly enough to uncover the optimal architecture. Using ImageNet as a test bed, computer vision researchers have systematized the process of neural architecture search.[11] Initial methods consumed too many computational resources to exhaustively cover the search space. Inspired by the success of hand-designed architectures with recurring architecture motifs, such as ResNet36 and Inception35, later methods defined architectures with recurring cell structures and restricted the search space to designing this recurring cell.[12]

The ImageNet Challenge ended once the accuracy of its best models reached superhuman levels, at 97.3 percent. (Human accuracy on this data was about 95 percent.)[13] Other researchers have continued making incremental advancements, however, using the ImageNet data set to track their progress, and error rates have continued to fall, though certainly not as fast as in the first few years after the introduction of ImageNet. The error rate of the best model today is only 1.2 percent, down from 33.6 percent when the competition began back in 2009.[14]

These days, thanks to high accuracy and reasonable computing costs, object recognition is in wide use. Whenever you search for images on the Internet, you use the kinds of algorithms first developed for the ImageNet Challenge; the same goes for when your smartphone automatically groups your photos based on

whose face appears in the photo. Those are exactly the uses we had in mind when we first chose object recognition as our north star. But uses of object recognition go beyond that, from tracking players in sports to helping self-driving cars detect other vehicles.

Learning to recognize objects is only one form of learning to see, which is why computer vision (or visual intelligence) is a much broader field than object recognition. But there are important similarities between object recognition and other tasks in computer vision, such as object detection and activity recognition. Such similarities mean that a computer should not need to tackle a new task from scratch. In theory, a computer should be able to take advantage of the similarities, applying what it has learned from one task to perform a somewhat different task. For both computers and humans, this process of generalizing knowledge from one task to a similar one is called *transfer learning*.[15]

Humans are very good at transfer learning: once we know French, for example, it is not as hard to learn Spanish. And if you learned to read English as a child, that was certainly easier if you already knew how to speak English than if the language was entirely new to you. In fact, the ability to pick up on similarities between tasks, and to parlay this shared knowledge to help us learn new tasks, is one of the hallmarks of human intelligence.

Transfer learning can be tremendously helpful for AI, too, but it does not come naturally to computers; instead, we humans have to teach them. The way to help computers with transfer learning is through pretraining. The idea is that before you give a machine learning model a new challenge, you first train it to do something similar, using training data that are already known to be effective. In computer vision, that starting point is the object-recognition data in ImageNet. Once a new model gets trained through ImageNet, it should have a leg up on tackling a new kind of challenge. If this approach works, as we thought it would, then we have all the more reason to think that object recognition is a north star for visual intelligence.

That was the thinking behind our extension of the ImageNet Challenge to the problem of object detection. Object detection means recognizing an object in an image and specifying its location within the image. If you have ever seen a digital photograph of a group of people with a little rectangle drawn around each person's face, you have seen one application of object detection. Whereas the images in ImageNet contain just one object each, most real-world scenes include several objects, so object detection is a valuable extension of the kind of simple object recognition we had tested in the ImageNet Challenge.

Object detection had been an area of research before ImageNet, too, but the most common approach then was to first identify the areas within the image where an object (such as an animal) was likely to be, and then to focus on that area

and try to recognize that object (as a tiger, for example).[16] Once ImageNet became available, that second step became much easier.

Object detection has come a long way since then, with special-purpose detectors for different kinds of applications, such as self-driving cars, which need to be alert to other cars on the road.[17] Such advances beyond object recognition would not have been possible without the use of ImageNet to enable transfer learning.

But object detection was just a first attempt to apply ImageNet data to uses beyond object recognition. These days, for better or for worse, almost every computer vision method uses models pretrained on ImageNet.

None of that is to say that ImageNet has been useful for every computer vision task. A prominent example is medical imaging.[18] Conceptually, the task of classifying a medical image (such as a screening mammogram) is not very different from the task of classifying a photograph taken with a phone camera (such as a snapshot of a family pet). Both tasks involve visual objects and category labels, so both could be performed by a properly trained machine. In fact, they have been. But the methods have not been exactly the same. For one thing, you cannot use the ImageNet data set to train a computer to detect tumors; it simply has no data for this specialized task. What is more, it is not feasible to use the same basic approach: the professional expertise required to create high-quality training data to help with medical diagnosis is scarce and expensive. Put another way, it is impossible to use Mechanical Turk to create a high-quality medical data set, both due to the requirement of specialized expertise as well as regulatory restrictions. So instead of using carefully labeled examples (the process of "supervised learning"), AI for medical imaging is usually based on "semi-supervised learning," whereby the machine learns to find meaningful patterns across images without many explicit labels.[19]

Computer vision certainly has practical applications beyond health, including environmental sustainability. Researchers are already using machine learning to analyze large volumes of satellite images to help governments assess changes in crop yields, water levels, deforestation, and wildfires, and to track longitudinal climate change.[20] Computer vision can be helpful in education, too: when students are trying to learn to read bar charts or to study visual subjects like geometry and physics, computers that understand images have the potential to supplement the efforts of human teachers. Assistive technology could also help teachers generate content-appropriate quizzes.[21]

The use of ImageNet to generalize beyond object recognition also led to the discovery of a thorny problem for deep learning models: "adversarial examples," which are images that fool an AI into making blatant errors classifying an object.[22] A miniscule, humanly imperceptible tweak to a picture (sometimes even a single pixel!) can cause a model trained on ImageNet to mislabel it entirely.[23] An image

of a panda can thus get misclassified as a bathtub. Some kinds of errors are easier to understand as the result of spurious correlations: wolves are often photographed in snow, so a model that learns to associate snow with wolves could come to assume that the label "wolf" refers to "snow." It turns out that all models that use deep learning are vulnerable to attacks from adversarial examples, a fact that has spurred some researchers to work on ways to "vaccinate" training data against these attacks.

The problem of adversarial examples has also led the computer vision community to shift from a singular focus on accuracy. Although accuracy in object recognition certainly remains important, researchers have come to appreciate the value of other criteria for evaluating a machine learning model, particularly interpretability (which refers to the ability of a model to generate predictable or understandable inference results for human beings) and explainability (the ability of a model to provide post hoc explanations for existing black box models).[24]

The success of ImageNet has also prompted the computer vision community to start asking what data the next generation of models should be pretrained on. As an alternative to the expensive, carefully annotated, and thoroughly verified process used to create ImageNet, researchers have collected data from social media and scraped images with their associated text off the Internet.[25] Pretraining models from this "raw" data have opened up the possibility of "zero-shot adaptation," the process through which computers can learn without any explicit labels. In fact, models trained on such raw data now perform as well as models trained using ImageNet.[26]

Finally, the wide influence of ImageNet has opened the data set up to criticism, raising valid concerns we were not sufficiently attuned to when we began. The most serious of these is the issue of fairness in images of people.[27] For one thing, although we certainly knew early on to filter out blatantly derogatory image labels such as racial or gender slurs, we were not sensitive to more subtle problems, such as labels that are not inherently derogatory but could cause offense when applied inappropriately (such as labeling people based on clues to their religion or sexual orientation). In addition, certain concepts related to people are hard to represent visually without resorting to stereotypes, so attempts to associate images with these concept labels ("philanthropist" or "Bahamian," for example) perpetuate biases. Most Bahamian wear distinctive garb only on special, ceremonial occasions, but an image search for "Bahamian" based on ImageNet data would give a disproportionate number of such stereotypical images of people from the Bahamas. Another source of bias in search results is the inadequate diversity in the ImageNet data set, a bias that tends to get amplified during the manual cleanup stage, when human annotators resort to racial and gender stereotypes in their labeling. Women and ethnic minorities are already underrepresented among real-world bankers, for example, but they are even *more* underrepresented in images

labeled as "banker." Although these problems of fairness are difficult to eliminate entirely, we have made research strides to mitigate them.[28]

The development of these new data sets has led to the need for a metabenchmark: a single evaluation scheme for multiple individual benchmarks (or a benchmark for comparing benchmarks). Without a metabenchmark, it is impossible to compare the performance of different machine learning models across different tasks and using different data sets.

In fact, one thing that has emerged is a lively debate about benchmarks themselves.[29] One side of the debate posits that the constant emergence of new benchmarks is a good sign, suggesting continued progress on north stars. On the other side is a concern that benchmarks encourage something akin to teaching to the test: the concern that what emerges from benchmarking are not superior models but models that optimize for high performance on an inherently imperfect benchmark.

Another serious concern is that a widely adopted benchmark amplifies the real-world effects of any flaws in the benchmark. There is a growing body of research, for example, on how benchmarks can perpetuate structural societal biases,[30] benefiting groups that are already dominant (particularly White males) while discriminating against marginalized groups (such as Muslims and dark-skinned females).[31]

In response to these concerns, pioneers in the field are radically rethinking benchmarking. One suggestion has been for human judges to generate inputs for which models would fail, thus creating increasingly harder testing criteria as models improve.[32] Another idea is to demand that benchmarks measure not only accuracy (which encourages designing to the benchmark) but also assess and reward progress on other valuable criteria, including bias detection.[33]

Where do we go next in computer vision? Other north stars beckon. One of the biggest is in the area of embodied AI: robotics for tasks such as navigation, manipulation, and instruction following. That does not necessarily mean creating humanoid robots that nod their heads and walk on two legs; any tangible and intelligent machine that moves through space is a form of embodied AI, whether it is a self-driving car, a robot vacuum, or a robotic arm in the factory. And just as ImageNet aimed at representing a broad and diverse range of real-world images, research in embodied AI needs to tackle the complex diversity of human tasks, from folding laundry to exploring a new city.[34]

Another north star is visual reasoning: understanding, for example, the three-dimensional relationships in a two-dimensional scene. Think of the visual reasoning needed to follow even the seemingly simple instruction to bring back the metal mug to the left of the cereal bowl. Following such instructions certainly requires more than vision, but vision is an essential component.[35]

Understanding people in a scene, including social relationships and human intentions, adds yet another level of complexity, and such basic social intelligence is another north star in computer vision.[36] Even a five-year-old can guess, for example, that if a woman is cuddling with a little girl on her lap, the two people are very likely mother and daughter, and that if a man opens a refrigerator, he is probably hungry; but computers do not yet have enough intelligence to infer such things. Computer vision, like human vision, is not just perception; it is deeply cognitive.

There is no question that all these north stars are huge challenges, bigger than ImageNet ever was. It is one thing to review photos to try to identify dogs or chairs, and it is another to think about and navigate the infinite world of people and space. But it is a set of challenges well worth pursuing: as computers' visual intelligence unfolds, the world can become a better place. Doctors and nurses will have extra pairs of tireless eyes to help them diagnose and treat patients. Cars will run more safely. Robots will help humans brave disaster zones to save the trapped and wounded. And scientists, with help from machines that can see what humans cannot, will discover new species, better materials, and uncharted frontiers.

AUTHORS' NOTE

The authors are grateful for the support of the Office of Naval Research MURI grants, a Brown Institute grant, the Stanford Institute for Human-Centered Artificial Intelligence, and the Toyota Research Institute.

ABOUT THE AUTHORS

Li Fei-Fei, a Fellow of the American Academy since 2021, is the Sequoia Capital Professor and Denning Family Co-Director of the Stanford Institute for Human-Centered Artificial Intelligence. She is an elected member of the National Academy of Engineering and the National Academy of Medicine. She has published in such journals as *Nature*, *Proceedings of the National Academy of Sciences*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Robotics Research*, *International Journal of Computer Vision*, and *The New England Journal of Medicine*.

Ranjay Krishna is an Assistant Professor at the Allen School of Computer Science & Engineering at the University of Washington. He has published essays in such journals as *International Journal of Computer Vision* and academic book chapters with Springer Science-Business Media, and has presented academic conference papers at top-tier computing venues for computer vision, natural language processing, and human-computer interaction.

ENDNOTES

[1] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59 (236) (1950): 433–460.

[2] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *AI Magazine*, Winter 2006.

[3] Andrew Parker, *In the Blink of an Eye: How Vision Started the Big Bang of Evolution* (London: Simon & Schuster, 2003), 316.

[4] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun, "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception," *Journal of Neuroscience* 17 (11) (1997): 4302–4311; and Russell Epstein and Nancy Kanwisher, "A Cortical Representation of the Local Visual Environment," *Nature* 392 (6676) (1998): 598–601.

[5] Jia Deng, Wei Dong, Richard Socher, et al., "ImageNet: A Large-Scale Hierarchical Image Database," in 2009 *IEEE Conference on Computer Vision and Pattern Recognition* (Red Hook, N.Y.: Curran Associates, Inc., 2009), https://image-net.org/static_files/papers/imagenet_cvpr09.pdf.

[6] Irving Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review* 94 (2) (1987), https://psycnet.apa.org/record/1987-20898-001.

[7] George A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM* 38 (11) (1995): 39–41.

[8] Jia Deng, Olga Russakovsky, Jonathan Krause, et al., "Scalable Multi-Label Annotation," in *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2014), 3099–3102.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* (2012): 1097–1105.

[10] Ibid.; Christian Szegedy, Wei Liu, Yangqing Jia, et al., "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2015), 1–9; Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," presented at the 2015 International Conference on Learning Representations, San Diego, California, May 7, 2015; Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2016); Saining Xie, Ross Girshick, Piotr Dollár, et al., "Aggregated Residual Transformations for Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017); and Chenxi Liu, Barret Zoph, Maxim Neumann, et al., "Progressive Neural Architecture Search," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2018).

[11] Barret Zoph and Quoc V. Le, "Neural Architecture Search with Reinforcement Learning," presented at the 5th International Conference on Learning Representations, Toulon, France, April 26, 2017; and Hieu Pham, Melody Guan, Barret Zoph, et al., "Efficient Neural Architecture Search via Parameters Sharing," *Proceedings of Machine Learning Research* 80 (2018): 4095–4104.

[12] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le, "Learning Transferable Architectures for Scalable Image Recognition," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2018), 8697–8710; and Hanxiao Liu, Karen Simonyan, Oriol Vinyals, et al., "Hierarchical Representations for Efficient Architecture Search," arXiv (2017), https://arxiv.org/abs/1711.00436.

[13] Olga Russakovsky, Jia Deng, Hao Su, et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115 (3) (2015): 211–252.

[14] Hieu Pham, Zihang Dai, Qizhe Xie, et al., "Meta Pseudo Labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2021), 11557–11568.

[15] Yusuf Aytar and Andrew Zisserman, "Tabula Rasa: Model Transfer for Object Category Detection," presented at the 2011 International Conference on Computer Vision, Barcelona, Spain, November 6–13, 2011; Maxime Oquab, Leon Bottou, Ivan Laptev, et al., "Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2014), 1717–1724; and Zhizhong Li and Derek Hoiem, "Learning without Forgetting," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2016).

[16] Shaoqing Ren, Kaiming He, Ross Girshick, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems* 28 (2015): 91–99; and David A. Forsyth and Jean Ponce, *Computer Vision: A Modern Approach* (Hoboken, N.J.: Prentice Hall, 2011).

[17] Bichen Wu, Forrest Iandola, Peter H. Jin, and Kurt Keutzer, "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017), 129–137.

[18] Varun Gulshan, Lily Peng, Marc Coram, et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *The Journal of the American Medical Association* 316 (22) (2016): 2402–2410; Geert Litjens, Thijs Kooi, and Babak Ehteshami Bejnordi, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis* 42 (2017): 60–88; and Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers, "Guest Editorial: Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," *IEEE Transactions on Medical Imaging* 35 (5) (2016): 1153–1159.

[19] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, "Semi-Supervised Learning," *IEEE Transactions on Neural Networks* 20 (3) (2009): 542–542; David Berthelot, Nicholas Carlini, Ian Goodfellow, et al., "Mixmatch: A Holistic Approach to Semi-Supervised Learning," *Advances in Neural Information Processing Systems* 32 (2019); and Kihyuk Sohn, David Berthelot, Chun-Liang Li, et al., "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence," *Advances in Neural Information Processing Systems* 33 (2020).

[20] Neal Jean, Marshall Burke, Michael Xie, et al., "Combining Satellite Imagery and Machine Learning to Predict Poverty," *Science* 353 (6301) (2016): 790–794.

[21] Chris Piech, Jonathan Spencer, Jonathan Huang, et al., "Deep Knowledge Tracing," *Advances in Neural Information Processing Systems* 28 (2015).

[22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al., "Intriguing Properties of Neural Networks," arXiv (2013), https://arxiv.org/abs/1312.6199.

[23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," presented at the 2015 International Conference on Learning Representations, San Diego, California, May 9, 2015; Szegedy et al., "Intriguing Properties of Neural Networks"; and Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," presented at the Sixth International Conference on Learning Representations, Vancouver, Canada, April 30, 2018.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: Association for Computing Machinery, 2016); Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1 (5) (2019): 206–215; and Ričards Marcinkevičs and Julia E. Vogt, "Interpretability and Explainability: A Machine Learning Zoo Mini-Tour," arXiv (2020), https://arxiv.org/abs/2012.01805.

[25] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, et al., "Exploring the Limits of Weakly Supervised Pretraining," in *Proceedings of the 2018 European Conference on Computer Vision* (New York: Computer Vision Foundation, 2018), 181–196; Ranjay Krishna, Yuke Zhu, Oliver Groth, et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision* 123 (2017): 32–73; and Tsung-Yi Lin, Michael Maire, Serge Belongie, et al., "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2014).

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning Transferable Visual Models from Natural Language Supervision," arXiv (2021), https://arxiv.org/abs/2103.00020; Zirui Wang, Jiahui Yu, Adams Wei Yu, et al., "Simvlm: Simple Visual Language Model Pretraining with Weak Supervision," arXiv (2021); and Chao Jia, Yinfei Yang, Ye Xia, et al., "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision," arXiv (2021), https://arxiv.org/abs/2102.05918.

[27] Chris Dulhanty and Alexander Wong, "Auditing ImageNet: Towards a Model-Driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets," arXiv (2019), https://arxiv.org/abs/1905.01347; Pierre Stock and Moustapha Cisse, "ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases," in *Proceedings of the European Conference on Computer Vision* (Cham, Switzerland: Springer Nature, 2018); and Eran Eidinger, Roee Enbar, and Tal Hassner, "Age and Gender Estimation of Unfiltered Faces," *IEEE Transactions on Information Forensics and Security* 9 (12) (2014): 2170–2179.

[28] Kaiyu Yang, Klint Qinami, Li Fei-Fei, et al., "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy," in *FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2020), 547–558.

[29] Sanjeev Arora and Yi Zhang, "Rip van Winkle's Razor: A Simple Estimate of Overfit to Test Data," arXiv (2021), https://arxiv.org/abs/2102.13189; and Avrim Blum and Moritz Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions," *Proceedings of Machine Learning Research* 37 (2015): 1006–1014.

[30] Antonio Torralba and Alexei A. Efros, "Unbiased Look at Dataset Bias," in *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2011), 1521–1528; Vinay Uday Prabhu and Abeba Birhane, "Large Image Datasets: A Pyrrhic Win for Computer Vision?" arXiv (2020), https://arxiv.org/abs/2006.16923; and Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356 (6334) (2017): 183–186, https:// science.sciencemag.org/content/356/6334/183.

[31] Abubakar Abid, Maheen Farooqi, and James Zou, "Persistent Anti-Muslim Bias in Large Language Models," arXiv (2021), https://arxiv.org/abs/2101.05783.

[32] Ozan Sener and Silvio Savarese, "Active Learning for Convolutional Neural Networks: A Core-Set Approach," presented at the Sixth International Conference on Learning Representations, Vancouver, Canada, May 1, 2018; and Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training Region-Based Object Detectors with Online Hard Example Mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2016).

[33] Moin Nadeem, Anna Bethke, and Siva Reddy, "Stereoset: Measuring Stereotypical Bias in Pretrained Language Models," arXiv (2020), https://arxiv.org/abs/2004.09456; and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91.

[34] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, et al., "BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments," presented at the Conference on Robot Learning, London, England, November 9, 2021; Eric Kolve, Roozbeh Mottaghi, Winson Han, et al., "AI2-THOR: An Interactive 3D Environment for Visual AI," arXiv (2017), https://arxiv.org/abs/1712.05474; and Xavier Puig, Kevin Ra, Marko Boben, et al., "VirtualHome: Simulating Household Activities via Programs," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2018).

[35] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, et al., "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017); Adam Santoro, David Raposo, David G. T. Barrett, et al., "A Simple Neural Network Module for Relational Reasoning," presented at the 31st Conference on Neural Information Processing Systems, Long Beach, California, December 6, 2017; and Justin Johnson, Bharath Hariharan, Laurens van der Maaten, et al., "Inferring and Executing Programs for Visual Reasoning," in *Proceedings of the 2017 IEEE International Conference on Computer Vision* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2017).

[36] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, et al., "Social LSTM: Human Trajectory Prediction in Crowded Spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2016); and Alina Kuznetsova, Hassan Rom, Neil Alldrin, et al., "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," *International Journal of Computer Vision* 128 (7) (2020).