Edited by Gillian Rose

Seeing the City Digitally

Processing Urban Space and Time

Amsterdam
University
Press

CITIESANDCULTURES

2. Deep Learning the City: The Spatial Imaginaries of AI

Joel McKim

Abstract
This chapter examines how deep learning neural networks and computer vision technologies are impacting the design, organization and occupation of cities. It begins by providing a brief history of the development of "deep learning" approaches to artificial intelligence. The chapter then focuses on the ways artists and designers have begun to engage with deep learning and computer vision in order to highlight critical questions, especially about the ethical issues surrounding the training datasets these systems depend on. The chapter discusses three art and design examples that shift focus specifically towards the city and spatial concerns, considering the ways these works explore machine learning (the opportunities it presents and the problems it raises) within a specifically architectural or urban context.

Keywords: deep learning, artificial intelligence, art, design, architecture

In the summer of 2019, the subterranean boiler room of New York's popular Chelsea Market opened to the public for the first time. Transformed from its original use, the room was now an art space, run by ARTECHOUSE, a self-described digital art organization dedicated to experiments in art and technology with exhibition venues in Washington and Miami, in addition to New York. For the inaugural exhibition in the 6,000 square foot boiler room, visitors were invited to enter a *Machine Hallucination* designed by Turkish-born artist Refik Anadol. Showing off the space's sophisticated projection technology, Anadol's immersive installation covered virtually every surface of the room in the kind of inexact, morphing images we've come to associate with artificial intelligence. Anadol's work is indeed an

Rose, G. (ed.), *Seeing the City Digitally. Processing Urban Space and Time*. Amsterdam: Amsterdam University Press, 2022
DOI 10.5117/9789463727037_CH02

experiment in AI-generated images – the artist has been working with AI and machine learning since completing a residency at Google's Artists and Machine Intelligence Program in 2016. The *Machine Hallucination* installation seems doubly relevant to a chapter on deep learning technologies and the city – the work is itself architectural, enveloping an interior space in a surround of AI-generated visuals, and those images, however dream-like or vague, are also recognizably urban. Anadol trained his machine-learning system on 100 million photographs of New York City found on social networks, effectively teaching it to produce its own images of the city based on this archive of public memories.

In some ways Anadol's work is representative of a growing number of artists and designers employing AI technologies such as machine learning in their work, sometimes as methods of aesthetic experimentation and other times as a means of questioning the social, political and economic impact of these fast-developing technologies. Understandably, the human form, and the human face in particular, has featured prominently in many of these works, with artists creating new AI-generated forms of portraiture or producing critical design projects examining the implications of machine-learning powered systems of facial recognition or human classification. Anadol's more unusual focus on images of the built environment invites a consideration of how these technologies are being deployed in the areas of architecture and city planning, but also how artists and designers are creating works that explore questions of AI and urban space. This chapter will outline some of the ways artists and designers are working with deep learning technologies, while highlighting works that address the images and spaces of the city specifically. At the risk of exhausting the limits of both my own technical knowledge and the patience of my readers, the chapter will begin by providing a brief history of the development of "deep learning" approaches to AI. While the computational and mathematical details of AI and machine learning systems can be difficult to summarize effectively or succinctly, I believe it's becoming increasingly important for scholars of the arts and humanities to attempt to engage with these systems at a technical level. As these technologies become central to contemporary visual culture, we need to develop a better understanding of the computational infrastructures that are producing a growing number of the cultural objects and images that surround us (from notorious "deep fake" videos, to AI "up-scaled" video games, to algorithmically filtered photographs).[1] After providing an

1 A number of arts and humanities-based research projects are beginning to take on the task of mapping the aesthetic and cultural significance of new developments in machine imaging

introduction to this technical context and terminology, the chapter will then outline some of the ways artists and designers have begun making use of deep learning technologies, highlighting the critical questions that have surfaced in this work. The ethical issues surrounding the training datasets these systems depend on emerges as a recurrent theme. Finally, the chapter will discuss three art and design examples that shift focus specifically towards the city and spatial concerns, considering the ways these works explore machine learning (the opportunities it presents and the problems it raises) within a specifically architectural or urban context, namely: the *Uncanny Rd.* online generative tool, Simone C. Niquille's CGI-based film *Homeschool*, and a trio of Forensic Architecture investigations: *Triple-Chaser*, *The Battle of Ilovaisk*, and *Model Zoo*.

A very brief history of deep learning

Deep learning (a term that now circulates frequently, but often without a great deal of explanation) is a specific approach to artificial intelligence and machine learning that involves a method based on a hierarchy of concepts. The fundamental idea being that a machine can learn more complex concepts by building on simpler concepts. As a result, the approach usually involves the use of multi-layered, and therefore "deep", artificial neural networks. We could imagine, for instance, a neural network trained to recognize hand-written numbers, a frequent example used in introductions to machine learning (see Nielsen 2019 and Bishop 2006). Early layers of the network might recognize very simple forms like edges, feeding this information forward to subsequent layers that recognize increasingly complex patterns (like loops or intersecting lines), until an output layer eventually recognizes the form of the numbers themselves. As Goodfellow, Bengio and Courville outline in their 2016 textbook on the subject, deep learning is a solution to the problem of machine learning that has a long history with an ebb and flow of acceptance within the field of AI research. They date the emergence of the concept of deep learning as far back as the 1940s, with its current resurgence as the dominant paradigm of AI beginning in 2006 (12). Three waves of development during this quite long history are identified by the

and computer vision, including the Machine Vision in Everyday Life research project led by Jill Walker Rettberg at the University of Bergen, the Operational Images and Visual Culture project led by Jussi Parikka at FAMU in Prague, and my own Pre-Histories of Machine Vision research project conducted at the V&A Museum.

authors: a cybernetic moment in the 1940s through 1960s that eventually wanes; a return to the concept through notions of "connectionism" in the 1980s and 1990s; and the current period spurred on by the breakthroughs of contemporary computer scientists like Geoffrey Hinton. I'll attempt to provide here a very rough sketch of the development of deep learning across these three waves or periods.

The first cybernetic moments of deep learning research emerged from early neural network research and an interest in models of biological brain function that were developing at the time. In 1943 the neuroscientist and cybernetician Warren McCulloch and the logician Walter Pitts proposed the first computational model of a neural net comprised of individual, largely undifferentiated neurons (the basic working unit of the brain, processing and transmitting cellular signals). Inspired by the extremely influential ideas of information theory being formulated by both Claude Shannon and Norbert Wiener at the time, McCulloch and Pitts proposed that the biological system of information exchange that is the nervous system could find analogous form in the logic processing of mathematics. We can view this as the beginnings of a long tradition of conceiving of the human brain as essentially a computation machine and therefore comparable to the digital computers just beginning to emerge at the time. McCulloch and Pitts begin their 1943 paper "A Logical Calculus of the Ideas Immanent in Nervous Activity" with the claim, "Because of the 'all-or-none' character of nervous activity, neural events and the relations among them can be treated by means of propositional logic" (McCulloch & Pitts 1943). In other words, the McCulloch-Pitts neuron, the basic unit of their model, was conceived as kind of logic gate – a linear mathematical function capable of taking a series of weighted inputs and aggregating them to produce a single output or decision. This essential premise – that an artificial neural network is made of a network of connected neurons, each one a mathematical function processing inputs according to varying weights – remains the foundation of contemporary deep learning.

The McCulloch-Pitts neuron would become the inspiration point for artificial neurons to follow, most notably the perceptron algorithm produced by the psychologist Frank Rosenblatt in 1958. Rosenblatt developed the perceptron at the Cornell Aeronautical Laboratory, funded by the US Office of Naval Research (ONR). Although first implemented as software running on an IBM mainframe computer, Rosenblatt intended for the perceptron to be realized as a custom-built machine, a goal which eventually materialized in the form of the "Mark I Perceptron" in the early 1960s. Image recognition was a central task for neural networks from the outset and the first use of

the perceptron involved connecting the machine to a simple camera system in which a lighted object was registered by a 20 × 20 array of cadmium sulphide photocells, producing a primitive 400 pixel image (Bishop 2006, 196). The photocells were wired to the neurons of the perceptron at random, demonstrating the system's ability to learn independently. An important distinguishing point from the McCulloch-Pitts neuron was the perceptron's ability to adjust the weighted values of the inputs automatically, rather than by human operator. Rosenblatt's research generated considerable public attention, but he was also considered to be prone to overclaiming, issuing "steady and extravagant statements about the performance of his machine" (McCorduck 2004, 105). After listening to Rosenblatt's initial 1958 press conference for the perceptron, *The New York Times* gushed: "The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of existence" (in Olazaran 1996 621).

Marvin Minsky was one important figure in the AI community irritated by Rosenblatt's bluster. The two scientists had attended the same high school in the Bronx and had maintained a rivalry throughout their careers (McCorduck 2004, 106). Minsky and Seymour Papert's 1968 book *Perceptrons: An Introduction to Computational Geometry* exposed some of the perceptron's limitations in relation to pattern recognition, classification, and its ability to internally represent its own act of perception. Minksy and Papert who favoured the rival "symbolic" approach to AI have more recently been accused of falsely characterizing the abilities of neural networks, focusing exclusively on the limitations of a single layer perceptron rather than the potential of multi-layered neural networks (which was already evident at the time). While suggesting the story is actually a more complicated one, Mikel Olazaran acknowledges, "according to the official history of the controversy, after Minsky and Papert's study, the neural-net approach was rejected and abandoned" (1996, 640).

According to Goodfellow, Bengio and Courville, the resurgence of interest in neural networks in the 1980s emerged out of the interdisciplinary field of cognitive science through a movement called "connectionism", which rekindled the notion that an interactive network of simple computational units was capable of generating intelligent behaviour (2016, 16). Many of the algorithms still in use in the machine learning of today were developed or optimized during this period. As Adrian Mackenzie notes in his book *Machine Learners*, "the algorithms such as back-propagation used in neural nets have not [...] been radically transformed in their core operations since the 1980s, and even then the algorithms (principally gradient descent)

were not new" (2017, 191). Put very simply, back-propagation is an algorithm by which a neural network is capable of optimizing the internal weights of the functions operating in its neurons, a key process in its ability to calibrate and learn. These algorithmic advances included breakthroughs in the field of computer vision and image processing, such as the development of convolutional neural networks – a class of deep learning network still considered to be the most effective for image recognition and classification (Fukushima 1980, LeCun et al. 1999). Convolutional neural networks were inspired by biological visual cortex systems and the sensory processing experiments of the neurophysiologists Hubel and Wiesel (1959). To again simplify greatly, a convolutional neural network employs operations of sub-sampling, filtering and synthesizing (a process of "convolving") in order to optimize its ability to recognize patterns in images.

Goodfellow, Bengio and Courville date the contemporary moment of deep learning to 2006 when significant breakthroughs in the effectiveness and efficiency of neural networks begin to emerge. Given that the basic premise of artificial neural networks and even some of the algorithms still in use date back to the mid-twentieth century, it seems fair to ask what brought about this relatively recent explosion of deep learning development. Most accounts of the growth of the field highlight two factors: the acceleration of computational processing power that has made feasible increasingly large or more efficient neural networks made of multiple layers (sometimes over a hundred), and the availability of large, often tagged, data sets used to train these neural networks. The availability of these large data sets has been fuelled in part by the expanded circulation and archiving of media online. Convolutional neural networks, for example, require images to learn and lots of them. The mass posting of photographs online that has occurred over the past two decades provides an ideal training resource for these networks.

An important example of these two factors coming together (increased processing power and the availability of large training sets) was the development of the convolutional neural network AlexNet, designed by Alex Krizhevsky at the University of Toronto, and published with Ilya Sutskever and Geoffrey Hinton (2012). AlexNet competed in the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) that has been a benchmark for computer vision developments. The competition called on research teams to use their deep learning neural networks to classify images from the ImageNet dataset. ImageNet, arguably the most significant computer vision training set, was first unveiled in 2009 by a team of AI researchers at Stanford and Princeton led by Professor Fei-Fei Li, who once described the project as an attempt "to map out the entire world of objects" (in Gershgorn 2017). The

dataset now consists of more than 14 million images, often scraped from online photo-sharing sites like Flickr and tagged by a crowdsourced army of workers into over 20,000 categories. These manually labelled datasets are often called "ground truth data" within discourses of deep learning (Schmidt 2019). AlexNet, employing an eight-layer convolutional neural network powered by two graphics processing units (GPUs), outperformed its competitors on its ability to correctly recognize or classify images in the ImageNet collection (images ranging from "container ships" to "Siamese cats"), achieving an impressively low error rate (Wei 2019).

A recent deep learning breakthrough has led to neural networks capable of not only classifying images, but also creating them, a development that has probably played the largest role in bringing wider technical advances in deep learning to public attention. Ian Goodfellow and colleagues invented Generative Adversarial Networks (GANs) in 2014 while Goodfellow was a student at the Université de Montréal under the supervision of Yoshua Bengio (Goodfellow et al. 2014). GANs involve engaging two neural networks (trained on the same dataset) in a kind of recognition game. A "generative network" produces images intended to pass as "candidates" for the dataset. A second "discriminative network" evaluates these generated images, determining how likely they are to be "real" images from the set. Through the learning mechanism of backpropagation both the generative network and the discriminative network gradually become better at their roles in this computational game of fool or be fooled. GANs are largely responsible for producing what has become the popular visual culture of AI, helping to create everything from the infamous "deep fakes" circulating online (the disturbingly iconic Jennifer Lawrence/Steve Buscemi mash-up video, for example) to the AI-generated *Portrait of Edmond Belamy* produced by the Paris-based Obvious collective that caused a media stir when it sold at auction for $432,000 in the Autumn of 2018. It's GAN technology, incidentally, that powered the machine hallucinations of New York City displayed by Refik Anadol in the Chelsea Market boiler room in 2019.

Deep learning in art and design

The art and design projects that make use of deep learning technologies, and GANs most frequently, are often both exploratory and critical in nature. Through initiatives such as the website *This Person Does Not Exist*, even AI industry insiders like Philip Wang (a software engineer at Uber at the time of the site's creation) strike a cautionary note regarding the deceptive potential

of the technology. True to its name, *This Person Does Not Exist* generates extremely convincing photorealistic images of otherwise non-existent people. It does so by employing StyleGAN, a generative adversarial network designed by engineers at Nvidia, the leading producer of the GPUs that provide the processing power for most neural networks. Each refresh of the page produces yet another person that does not exist. In describing his motivation for creating the site Wang explains, "I just hope my demonstration raises awareness. Those who are unaware are most vulnerable to this technology" (in Paez 2019). Although despite this expression of concern, Wang also sees positive potential for deep learning technologies and accompanies each image with the tag line "Don't panic. Learn how it works", along with links to YouTube videos explaining the technical details of the GAN-powered human face synthesis algorithms at work in creating these images.[2]

Many of the art and design projects exploring the growth of deep learning and computer vision technologies question the role played by the training sets and hierarchies of classification that serve as the underlying infrastructure of these systems. Adam Harvey's ongoing *MegaPixels* initiative is a good example of this critical perspective, a project described by the artist and researcher as an investigation of "the ethics, origins, and individual privacy implications of face recognition image datasets and their role in the expansion of biometric surveillance technologies" (Harvey 2019). Harvey questions the political implications of datasets such as MegaFace, a training set of 4,753,320 faces derived from public Flickr photo albums, analysing the metadata connected to the dataset's images and revealing the potential violation of Creative Commons licenses involved in their use.

As mentioned in our short history of deep learning, ImageNet is the training set that has almost certainly contributed to recent developments in computer vision and machine learning more than any other. Unsurprisingly, the influential dataset has also been the focus of a number of critical art and design projects. The artist Trevor Paglen has placed ImageNet at the centre of two recent projects, his 2019 Barbican exhibition featured a newly commissioned work entitled *From "Apple" to "Anamoly"* – an array

2 The synthesized faces generated by contemporary GAN technology reproduce some of the desires and anxieties provoked by earlier iterations of computational technologies. These faces recall *Time* magazine's controversial cover for its special fall 1993 issue on immigration and multiculturalism, "The New Face of America", featuring the image of a woman purportedly produced by morphing together the facial features of multiple racial and ethnic groups. Donna Haraway critiques the eliding of messy biological and political difference represented by this technologically composited, universal "SimEve" (1995). We might question what new technophilic fantasies of identity accompany the endless non-faces produced by neural networks.

of approximately 30,000 printed photographs, all images derived from the training set, that virtually covered the sweeping wall of the Curve gallery. Paglen's work displays images belonging to a cross-section of ImageNet's categories, beginning rather innocently with images clustered around the labels "apple", "apple tree" and "fruit", before moving on to more complex and contentious examples, from "minibar" to "abattoir" to "divorce lawyer". As I wrote in a recent review of the exhibition: "It's when people first appear among the photographs of objects that we begin to realize how strange and troubling this exercise in image classification really is. The category 'picker' includes smiling recreational strawberry pickers alongside Indian tea-leaf pickers and impoverished children picking through waste in a landfill. Any contextual distinction between these images is apparently flattened out in the eyes of the machine" (McKim 2019). The biases and absurdities of ImageNet's structure of classification become obvious when we notice, for example, that the labels "investor", "entrepreneur" and "venture capitalist' present almost exclusively images of white, middle-aged men, whereas less flattering categories such as "selfish person", "moneygrubber" and "convict" are considerably more diverse.

As the artist and programmer Nicholas Malevé has pointed out, the classification system for ImageNet is reliant on the WordNet database of semantic relations developed at Princeton: "Pressing into service an existing classification system however brings in its own share of problems, omissions and decision-making issues. WordNet for instance unreflexively integrates and naturalizes racial and gender binaries and its structure contributes to reifying social norms" (2019). The potential problems associated with ImageNet's system of classification were further highlighted in Paglen's *ImageNet Roulette*, a project featured in the "Training Humans" exhibition at the Osservatorio Fondazione Prada that Paglen and Kate Crawford co-curated in 2019. *ImageNet Roulette* is a computer vision system that captures the video image of gallery visitors and assigns them labels from the ImageNet's people categories (an online version of the work was also made available). The labels are often uncomplimentary, gendered and even racist, which Paglen and Crawford defend as a provocation to question the inherent prejudices of the ImageNet dataset and these forms of human categorisation more generally. In their "Excavating AI" text accompanying the exhibition they write: "ImageNet is an object lesson, if you will, in what happens when people are categorized as objects" (2019).

London's Photographers' Gallery has also thoroughly and provocatively explored the politics and ethics of image training sets in their year-long programme of events and commissions entitled "Data / Set / Match", led by curators

Katrina Sluis and Jon Uriarte and running over 2019-2020. The programme included exhibiting 14,197,122 photographs from ImageNet on the gallery's Media Wall, the images cycling through at a rate of ninety milliseconds per image following a computer script written by Malevé. *The Future Is Here!,* a video work by Mimi Onuoha commissioned by the gallery, explores the exploitation of labour involved in the annotation of training sets, a process often involving a dispersed group of crowdsourced workers connected through micro-tasking platforms such as Amazon's Mechanical Turk. Onuoha's video depicts the otherwise unseen domestic working spaces of these poorly compensated image taggers, many based in Venezuela. As Florian A. Schmidt describes in his response to Onuoha's video, "they work as freelance sub-sub-contractors, switching back and forth between different platforms that funnel the work from supranational corporations to people in the Global South" (2020). The artist Anna Ridler (who also featured in the "Data / Set /Match" programme) has likewise confronted the problematic ethics of training sets, both in terms of the labour practices involved in their creation and the classification systems they draw on. In works such as *Fall of the House of Usher* and *Mosaic Virus*, Ridler insists on producing her own datasets, employing machine learning systems trained on thousands of images she painstakingly creates herself. For the *Mosaic Virus* project, for example, Riddler photographed and hand classified over ten thousand tulips acquired during a single tulip season in Amsterdam.

Deep learning and the city

The critical attention focused on deep learning technologies by artists, designers and curators, in recent years in particular, has done much to expose the complex processes and infrastructures that underpin the purported AI revolution now underway. Understandably, many of these projects have placed the human at the centre of their investigations – questioning the systems of categorization, surveillance and deception machine learning may engender, as well as the precarious labour practices that enable their creation. And the role played by image training sets, the often-unseen foundations or 'ground truth' of deep learning, has justifiably attracted particular scrutiny. The ways in which computer vision and machine learning technologies are transforming urban space may have received comparably less attention from artists and designers, but the questions of automation, surveillance and classification that these projects address are of course also deeply connected to spatial concerns. As Shannon Mattern aptly states in her examination of the growth of intelligent mapping technologies, "with the stakes so high,

we need to keep asking critical questions about how machines conceptualize and operationalize space. How do they render our world measurable, navigable, usable, conservable?" (2017). With this call in mind, the final section of this chapter will outline three recent art and design projects that do take as their primary focus the built environment and architectural or urban space, namely: the *Uncanny Rd.* online generative tool, Simone C. Niquille's CGI-based film *Homeschool*, and a trio of Forensic Architecture investigations: *Triple-Chaser*, *The Battle of Ilovaisk*, and *Model Zoo*.

Uncanny Rd. is a web tool designed by software developers Anastasis Germanidis and Cristóbal Valenzuela, the co-founders of RunwayML, a popular machine learning programme aimed at artists and designers. The project involves a relatively simple interface that provides users with a coloured map of a street scene which can be populated, according to preference, with a number of different object labels, such as streetlamps, pedestrians, cars, etc. This "semantic map" showing only the basic outline of objects within the scene is synthesized by a GAN trained on city streets, generating a somewhat distorted or impressionistic image of a streetscape with a slightly post-apocalyptic aesthetic – something reminiscent of *Mad Max* or the *Borderlands* videogame franchise. The project is described on the site itself as: "Collectively hallucinating a never-ending road using Generative Adversarial Neural Networks." Apart from being an amusing interactive drawing tool that showcases some of the generative capabilities of GANs, *Uncanny Rd.* is perhaps more significant for drawing attention to the training set it relies on, the Cityscapes Dataset.[3] Produced by the Max Plank Institute, TU Darmstadt, and Daimler AG R&D (the research arm of Mercedes-Benz), Cityscapes is an annotated or labelled dataset of recorded stereo video sequences captured in streets from fifty cities, mostly located in Germany. A Mercedes hood ornament appears at the bottom of every image produced by the *Uncanny Rd.* site, a giveaway as to the origins of the neural network's training material.

The motivation for producing Cityscapes, clearly not to enable the creation of playful online drawing tools, is made quite explicit in an accompanying research paper describing the dataset as "specifically tailored for autonomous driving in an urban environment" (Cordts et al. 2016, 1-2). To this end, Cityscapes provides "semantic urban scene understanding", or put more simply, it identifies and categorizes objects that appear in its large video collection of street scenes. The "semantic" object labels *Uncanny Rd.*

3 My thanks to Bernd Behr for sharing his insights on the significance of the Cityscapes Dataset.

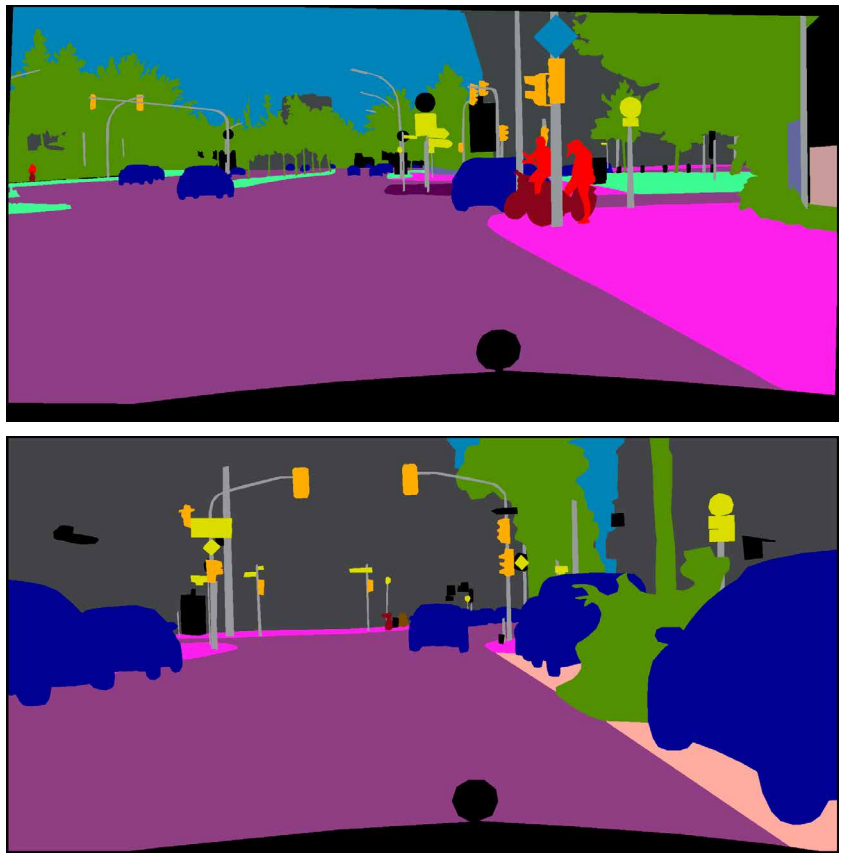

Figure 2.1. Semantic maps from the Cityscapes Dataset (Hamburg and Dusseldorf).

makes available to its users are pulled directly from the "class definitions" established in the Cityscapes Dataset, categories ranging from sidewalk, to bicycle, to person, to guard rail – all things that might be very useful for an autonomous vehicle to be able to recognize with a high degree of accuracy. The semantic mapping of Cityscapes is thus one component (along with detection technologies like Lidar) of the complex "sense-making capacities" of autonomous vehicles carefully considered by Sam Hind in his chapter in this volume. While many of the datasets used in the computer vision research of autonomous vehicle companies are proprietary, Cityscapes has had a wider general influence due to its public availability. It surfaces in a number of additional research areas, for example, in the video-to-video synthesis work conducted by Ting-Chun Wang and others at Nvidia and MIT, research that takes *Uncanny Rd.* a step further by generating photo-realistic moving video from the semantic maps that Cityscapes enables (Wang et al. 2018).

While perhaps not as obviously problematic as the issues of racial and gender bias inherent in a dataset like ImageNet, urban training sets such as Cityscapes nevertheless raise related questions of classification and standardization. Is it important, for example, that this widely influential dataset is based exclusively on scenes from German cities? What unintended consequences might arise from the public reliance on a Daimler AG produced training set, beyond the branding effect of the omnipresent Mercedes logo in every image generated from Cityscapes image data? What influence will a classification system attuned to the specific goals of autonomous vehicle design have on other forms of urban research making use of the dataset? Fiona McDermott articulates some of these concerns in her thoughtful work on the kinds of sensorial regimes produced by autonomous vehicle development. She writes that autonomous vehicles, "are only possible given huge amounts of collected and processed data, which begs the question as to how these exhaustive amounts of information might in turn have implications for the design and use of the space" (2019, 252). McDermott references the cautionary analysis of Florian Cramer who finds it all too easy to imagine an urban environment designed to be optimized for the limited category recognition of our current machine vision systems: "[A] ll cars and highways could be redesigned and rebuilt in such a way as to make them failure-proof for computer vision and autopilots. For example, by painting all cars in the same specific colors, and with computer-readable barcode identifiers on all four sides, designing their bodies within tightly predefined shaper parameters to eliminate the risk of confusion with other objects." (in McDermott 2019, 252).

What's clear from examples like Cityscapes Dataset is that computer vision technologies and the neural networks they rely upon are not only producing new machinic readings of the city, they are also altering the way humans view and interpret their urban surroundings. For Steve F. Anderson the current task is not to reinforce an opposition between organic human seeing and machine vision, given how inevitably intertwined the two have become, but instead to reflect on the ways human vision has been "reconstituted in dialogue with the computational" (2017, 82). However alien or uncanny the semantic maps or GAN-produced images of machine vision may appear, the forms of information they prioritize and the particular ways in which they segment and order the world shapes, for better or for worse, our own patterns of seeing and urban understanding. As the media philosopher Vilém Flusser noted of the computational images emerging in the 1970s and 80s, our technical images don't simply represent the outside world, they also envision or inform it: "Technical images are not mirrors

but projectors" (2011, 51). The images used and produced by deep learning networks constitute some of the most important technical images of our current age and they undoubtedly project a specific regime of computational vision on the contemporary city.

The precarious networks of labour involved in other processes of image classification are also very much present within the computer vision research of the automotive industry. In fact, Schmidt's research on the human workers teaching self-driving cars "to see" reveals the emergence of a new sector of specialist platforms catering specifically to the labour demands of deep learning dependant industries like autonomous vehicles. He notes, "probably the most important lesson from studying the crowdsourced production of AI training data is that in the relatively short time of one and a half years the automotive industry was able to access hundreds of thousands of new workers, through a labour supply chain of venture capital funded platforms which sprung up like mushrooms to cater for this new demand" (Schmitt 2019, 25). This dispersed network of urban workers, predominantly from the global south, is a less frequently acknowledged geographic by-product of this developing technology.

The impact of deep learning and machine vision on design and automation is being played out on multiple urban scales, ranging from the metropolitan to the domestic. The recent work of designer Simone C. Niquille moves us from a concern with autonomous mobility in the city to a consideration of the technologies of automation targeting interior space. Her animated film *Homeschool* (2019) exposes yet another image dataset, this time one used in the computer vision training of domestic robots. The film is set within the CGI interior of a home populated with rendered objects derived from SceneNet RGB-D, a training set produced by the Dyson Robotics Lab at Imperial College. In this case the dataset is comprised of computer generated or "synthetic" images rather than photographs or videos, as this presents a more effective way of producing the mundane scenes of domestic clutter that an automated vacuum cleaner, for example, might rely on in order to learn how to navigate its environment. After all, we don't tend to offer up photographs of our messy living rooms on Flickr, or at least not in the vast quantities required for deep learning.

Niquille's film was originally titled *Regarding the Pain of Spotmini*, referencing the smaller iteration of the dog-like Spot robot produced by Boston Dynamics, this miniature version being small and nimble enough to handle the confined spaces of domestic and office interiors. Using a method that can appear a little surreal, SceneNet RGB-D produces its database of images by allowing synthetic objects to randomly drop from the ceiling of a CGI room,

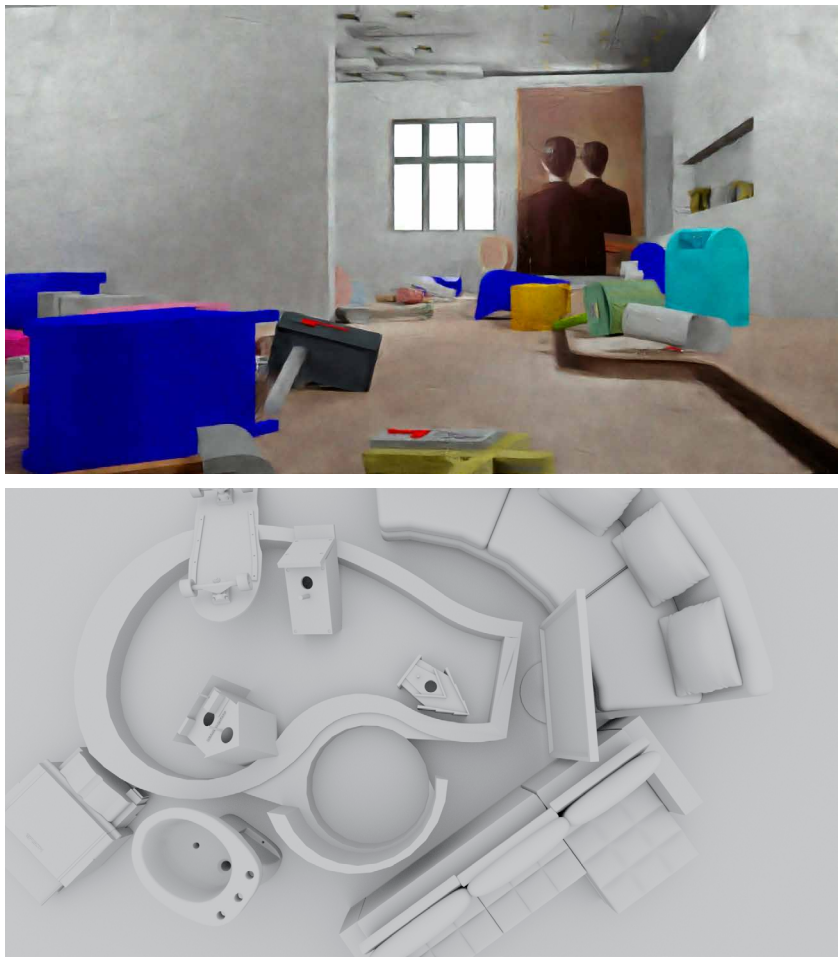

Figure 2.2. *Homeschool* (2019) by Simone C. Niquille. Courtesy of Simone C. Niquille.

settling according to the gravitational logic of a physics engine. Niquille's film presents the viewer with the anthropomorphized inner monologue of a robotic computer vision system as it "learns what a home is". The robotic protagonist moves about the space becoming gradually more proficient at naming objects like doors, plants and furniture. In a humorous, but also slightly sinister moment, the vision system approaches a CGI handgun lying on the floor of a living room that also contains a dining table and a child's pram. "Decoration? Toothbrush? Candle?" the voice asks, apparently struggling to identify the synthetic object. As the voice self-reflexively comments at the conclusion of the film: "The limits of my categories mean the limits of my world."

Niquille's interest lies in exploring the numerous decisions and assumptions of language that underpin something as apparently straightforward as the description and categorization of household objects. What are the logic parameters of what constitutes a chair in the eyes of a computer vision system? A piece of furniture with four legs? Anything we can sit down on? She explains, "Autonomous machines' computer-vision capabilities depend on the resolution of their training database. The database, however, is a subjective collection created by engineers, technicians or academic researchers. Once filtered through computer vision, this subjectivity becomes obscured: the seeing technology is too easily mistaken as an impartial agent" (Niquille 2019, 90). The inevitable tendency towards standardization involved in these systems is also an important consideration for Niquille. In a kind of recursive loop of uniformity, she reveals that the rendered objects included in the SceneNet RGB-D training set are themselves largely derived from yet another image dataset, the "Dataset for IKEA 3D Models" produced by MIT in 2013. The ubiquity of IKEA furniture makes it an ideal test case for computer vision research. Just as Cramer foresees cars, highways and city spaces being adapted to the requirements of machine vision, we might just as easily imagine a future of interior design standardization conforming to the learning needs of domestic automation and robotics. The particular projected viewpoint of neural networks thus has the potential to influence the organization of the urban from the infrastructural to the architectural.

The final example considered in this chapter also involves the use of synthetic datasets, but this time turned from the restrictive sphere of domestic interiors to the more expansive terrain of international urban conflict. For the past decade Goldsmiths' Forensic Architecture (FA) research group, led by Eyal Weizman, have employed advanced visualization technologies like digital animation and simulation in their important investigations of human rights violations, political violence and issues of environmental justice (Weizman 2017, McKim 2017). The incorporation of deep learning and computer vision techniques into the group's research methods is a more recent development, one supported by the arrival of FA members like software developer Lachlan Kermode.

The first demonstration of these new approaches can be seen in the agency's *Triple-Chaser* film, FA's response to an invitation to participate in the controversial 2019 Whitney Biennial. The exhibition had already been boycotted by a number of invited artists, a protest against the involvement of Whitney board vice-chairman Warren B. Kanders, whose company the Safariland Group produced tear gas munitions used by US agents against migrants at the US-Mexico border in an incident on November 25, 2018. The FA film, narrated by the musician David Byrne, documents the group's

Figure 2.3. *Triple-Chaser* (2019) by Forensic Architecture/Praxis Films. Courtesy of Forensic Architecture/Praxis Films.

process of training a machine learning classifier to search for images of the "triple-chaser" tear gas grenades manufactured by Defense Technologies, a subsidiary of Safariland. Able to locate only a hundred images of the triple-chaser grenade online (far too few to serve as a functional training set), FA turned to generating a synthetic image data set as a method of training their machine learning system.[4] Based on video footage of triple-chasers provided to FA by artists and activists and specifications available

4 A detailed account of the group's use of synthetic images is available in the FA report "Synthetic Data Generation: Development of Data Classification Tools".

in product catalogues, the group was able to create a digital 3D model of the grenade which could then be inserted into various background images (both computer generated and photo-realistic) in order to build a sizeable training set. Some of these images were produced using a process not unlike the one used to generate the images in the SceneNet RGB-D dataset, dropping CGI triple-chaser grenades randomly into scenes in order to produce a large variety of possible configurations. Having trained their machine learning system to identify the triple-chaser, FA is now deploying the classifier to search for the grenades across online images and video repositories, such as YouTube. The list of places where the group has already identified the use of Safariland-produced grenades against civilians is already long and includes Turkey, Peru, Iraq, Yemen, Egypt and Palestine, amongst other countries.

The *Triple-Chaser* film was both a provocation to the Whitney and an opportunity for FA to prototype a new method of research. A synthetic image approach to machine learning has since been employed in at least two subsequent investigations. *The Battle of Ilovaisk* investigation, commissioned by the European Human Rights Advocacy Centre (EHRAC) and the Ukrainian Legal Advisory Group (ULAG), called on FA to gather and analyse available evidence of the presence of the Russian military in Eastern Ukraine during a battle in the summer of 2014 between pro-Russian separatists and the Ukrainian Armed Forces. FA again experimented with the use of a machine learning classifier to help automate the process of analysing a large amount of open source information. This time the machine learning system was trained to recognize Russian military vehicles, such as the T-72B3 tank. Once trained, the classifier could then be programmed to automatically scour video platforms like YouTube.

Finally, FA's *Model Zoo* initiative, undertaken in collaboration with Bellingcat and Amnesty International, is the ongoing development of an open-source library of 3D models of weapons and munitions, along with various classifiers trained to identify them. A possible shared resource for multiple human rights organizations, the *Model Zoo* project confronts some of the barriers of access to deep learning technologies faced by non-commercial institutions. As will by now be clear, the effectiveness of machine learning in any domain is largely dependent on the availability of suitable training sets, which are expensive to produce and limited by image attainability. As a result, the production of datasets has been heavily weighted towards applications with the potential for large economic payoffs such as autonomous vehicles or industrial robotics. The *Model Zoo* initiative by FA is an attempt to ensure that the potential of deep learning technologies is not limited to either commercial ventures, with often problematic labour consequences,

or even more troubling forms of control or surveillance. The group's forays into machine learning therefore echo FA's longer tradition of turning the advanced visualization technologies that are too often the exclusive domain of state powers and corporate interests towards a decidedly different agenda of human rights activism. While deep learning technologies are already shaping the built environment on multiple levels, Forensic Architecture's experiments introduce the potential for a productive machine vision intervention in urban conflict zones with substantial geo-political implications.

Conclusion

The projects outlined above provide at least an indication of how deep learning technologies are already impacting the design, organization and occupation of cities. These works provoke specifically urban or architectural questions, while also raising issues that are present across a wider field of art and design concerned with machine learning and AI. The critical projects of the past several years have done much to expose the inner working and inherent pitfalls of the training sets and computer vision systems employed in human oriented machine learning systems. In spatially oriented fields ranging from driverless vehicles to domestic robotics, we find equivalent problems of bias, classification, and automation. In her insightful book *Cloud Ethics* Louise Amoore asserts that the most pressing ethicopolitical questions arising from neural networks are less those related to the common fears of automation breaking free from human control and more those occasioned by "a machine learning that generates new limits and thresholds of what it means to be human" (2020, 65). The examples highlighted in this essay reframe this question slightly, compelling us to ask what it now means to be human in an urban environment increasingly shaped by machine vision.

Whether through detailing technical histories or producing creative investigations there remains work to be done to better comprehend and contend with technologies that are having an undeniably transformative impact on contemporary visual culture and urban life. The most promising of these projects are not only critiques, they are also efforts at greater understanding and explorations of alternative applications. Niquille's *Homeschool*, for example, literally gives voice to the machinic intelligences increasingly embedded within our domestic spaces, while the work of Forensic Architecture encourages us to challenge the current use of these emergent technologies by envisioning ways to deploy them towards different and unanticipated political ends.

References

Amoore, Louise. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press.

Anderson, Steve F. 2017. *Technologies of Vision: The War Between Data and Images*. Cambridge, Mass: MIT Press.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Convolutional Neural Networks. NIPS Proceedings.

Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. 2016. "The Cityscapes Dataset for Semantic Urban Scene Understanding." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

Crawford, Kate and Trevor Paglen. 2019. "Excavating AI: The Politics of Images in Machine Learning Training Sets." https://www.excavating.ai/

Flusser, Vilem. 2011. *Into the Universe of Technical Images*. Minneapolis: University of Minnesota Press.

Forensic Architecture. 2018. "Synthetic Data Generation: Development of Data Classification Tools." https://synthetic-datasets.sfo2.digitaloceanspaces.com/website/FA_Synthetic-Data-Generation_Report.pdf

Fukushima, Kunihiko. 1980. "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36 (4): 193–202.

Gershgorn, Dave. 2017. "The Data that Transformed AI Research — And Possibly the World." *Quartz* https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

Goodfellow, Ian, Yoshua Bengio and Aaron Courville. 2016. *Deep Learning*. Cambridge, Mass: MIT Press.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. 2014. "General Adversarial Nets." NIPS Proceedings.

Haraway, Donna. 1995. "Universal Donors in a Vampire Culture: It's All in the Family. Biological Kinship Categories in the Twentieth-Century United States." In *Uncommon Ground: Rethinking the Human Place in Nature*, edited by William Cronon, 321-377. New York: W. W. Norton.

Harvey, Adam. 2019. "MegaPixels: Face Recognition Training Datasets." https://ahprojects.com/megapixels/

Hubel, David H. and Torsten Wiesel. 1959. "Receptive Fields of Single Neurones in the Cat's Striate Cortex." *Journal of Physiology* 148 (3): 574–91.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey Hinton. 2012. "ImageNet Classification with Deep

LeCun, Yann, Patrick Haffner, Léon Bottou and Yoshua Bengio. 1999. "Object Recognition with Gradient-Based Learning." In *Lecture Notes in Computer Science, Vol 1681*. Berlin: Springer.

Mackenzie, Adrian. 2017. *Machine Learners: Archaeology of a Data Practice*. Cambridge, Mass: MIT Press.

Mattern, Shannon. 2017. "Mapping's Intelligent Agents." *Places Journal*. https://placesjournal.org/article/mappings-intelligent-agents/

McCulloch, Warren S. and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115-133.

McDermott, Fiona. 2019. "New Sensorial Vehicles: Navigating Critical Understandings of Autonomous Futures." In *Architecture and the Smart City*, edited by Sergio M. Figueiredo, Sukanya Krishnamurthy, Torsten Schroeder, 247-251. London: Routledge

Malevé, Nicholas. 2019. "An Introduction to Image Datasets." *Unthinking Photography*. https://unthinking.photography/articles/an-introduction-to-image-datasets

McCorduck, Pamela. 2004. *Machines Who Think*. Natick, Mass: A K Peters.

McKim, Joel. 2017. "Speculative Animation: Digital Projections of Urban Past and Future." *Animation* 12 (3): 287–305.

———. "Trevor Paglen Trains His Sights on the Rise of Machine Vision." *Apollo Magazine*. https://www.apollo-magazine.com/trevor-paglen-machine-vision/

Minsky, Marvin and Seymour A. Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Mass: MIT Press.

Nielsen, Michael. 2019. *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com/chap1.html.

Olazaran, Mikel. 1996. "A Sociological Study of the Official History of the Perceptrons Controversy." *Social Studies of Science* 26 (3): 611–659.

Niquille, Simone C. 2019. "Regarding the Pain of SpotMini: Or What a Robot's Struggle to Learn Reveals about the Built Environment." In *Machine Landscapes: Architectures of the Post-Anthropocene*, edited by Liam Young, 84-91. Oxford: Architectural Design.

Paez, Danny. 2019. "'This Person Does Not Exist' Creator Reveals His Site's Creepy Origin Story." *Inverse*. https://www.inverse.com/article/53414-this-person-does-not-exist-creator-interview

Schmidt, Florian A. 2019. "Crowdsourced Production of AI Training Data: How Human Workers Teach Self-Driving Cars to See." Dusseldorf: Hans-Böckler-Stiftung.

———. (2020). "Unevenly Distributed." *Unthinking Photography*. https://unthinking.photography/articles/unevenly-distributed

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz and Bryan Catanzaro. 2018. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

Wei, Jerry. 2019. "AlexNet: The Architecture that Challenged CNNs." *Medium*. https://towardsdatascience.com/alexnet-the-architecture-that-challenged-cnns-e406d5297951

Weizman, Eyal. 2017. *Forensic Architecture: Violence at the Threshold of Detectability*. New York: Zone Books.

About the author

Joel McKim is Senior Lecturer in Digital Media and Culture and the Director of the Vasari Research Centre for Art and Technology at Birkbeck, University of London. He is the author of *Architecture, Media, and Memory: Facing Complexity in Post-9/11 New York* (Bloomsbury 2018) and recently completed a visiting fellowship at the V&A Museum, working on the research project "A Pre-History of Machine Vision".