BJHS Themes (2023), **8**, 171–187
doi:10.1017/bjt.2023.6

RESEARCH ARTICLE

Five experimentations in computer vision: seeing (through) images from Large Scale Vision Datasets

Bruno Moreschi

Faculty of Architecture and Urbanism, University of São Paulo, Brazil
Email: brunomoreschi@gmail.com

Abstract

Using images from large-scale vision datasets (LSVDs), five practice-based studies – experimentations – were carried out to shed light on the visual content, replications of historical continuities, and precarious human labour behind computer vision. First, I focus my analysis on the dominant ideologies coming from a colonial mindset and modern taxonomy present in the visual content of the images. Then, in an exchange with microworkers, I highlight the decontextualized practices that these images undergo during their tagging and/or description, so that they become data for machine learning. Finally, using as reference two counterhegemonic initiatives from Latin America in the 1960s, I present a pedagogical experience constituting a dataset for computer vision based on works of art at a historical museum. The results offered by these experimentations serve to help speculate on more radical ways of seeing the world through machines.

> 'Olha-me de novo. Com menos altivez.
> E mais atento.'
>
> (Look at me again. With less haughtiness.
> And more attentive.)
> 'Dez chamamentos ao amigo', poem by Hilda Hilst (free translation)

Introduction

During 2021, I followed the work routine of Pedro, who has been working remotely on human intelligence tasks (HITs), microtasks that computers cannot do efficiently, since 2015.[1] He notified me of every service he performed related to images and sent me screenshots of his monitor. During the COVID-19 pandemic, Pedro and I saw the outside world through images of pizza (mostly pepperoni), happy (white) families, lots of cats and dogs, and people (thin and healthy) playing sports, among many other trivial scenes. Once images like these are organized, tagged and/or described by microworkers like Pedro,

1 The names of workers on the Amazon Mechanical Turk platform have been changed to preserve anonymity and protect them from possible punishment by Amazon and the requesters.

they become material for machine-learning training datasets. Used to train artificial intelligence (AI), they are the backstage content of computer vision technologies used for social media, facial recognition, driverless vehicles and drone vision.[2]

The image files are organized in folders with a huge number of images. Amongst the many such datasets available on the Internet, I am interested in some of the bigger machine-learning training datasets, such as large-scale vision datasets (LSVDs) and their known models: ImageNet, Visual Genome, Open Images, Flick-Faces-HQ and TinyImages (once widely used, TinyImages is no longer available thanks to an investigation carried out by Vinay Uday Prabhu and Abeba Birhane).[3] ImageNet is the most emblematic LSVD amongst this selection. With 14,197,122 images it achieved a data scale never before seen.[4] This list of datasets is based on a questionnaire distributed to programmers on platforms like Discord, and answered by a hundred professionals.[5] This article begins an investigation into *how* and *what* is 'seen' in these images.

To understand them beyond their visual content, I draw on computer science, the visual arts, education and history and develop empirical results using a methodology known as practice-based research. This enables researchers to incorporate results from their creative practice into their academic investigation, as 'creative work in itself is a form of research and generates detectable research outputs'.[6] Creative outputs derived from this type of methodology may include images, sounds, videos, websites, performances and exhibitions. Smith and Dean also observe that they appear in academic investigations through textual formalizations.[7] The creative output and its theoretical discussion in text form are not independent, but work together to address the research question.

The creative outputs used here come from projects carried out over the last three years: an interactive website, postcard exchanges, programming codes, images generated from algorithms and so on. I refer to these projects as 'experimentations', using the term as defined by Annette Markham and Gabriel Pereira. Unlike 'exploration', 'observation' or 'contemplation', experimentations come with the idea of testing. They need not be carried out by academic researchers or scientists but can be less formalized, produced by an artist or even a non-expert.[8] They are characterized by being guided by a curiosity about something that is difficult to understand at first glance – such as computer vision. For Markham and Pereira, 'the lens of experimentalism – in the stereotypical sense of a scientific laboratory – provides a useful mindset for conceptualising and enacting participatory and interventionist research that seeks to promote critical data literacy'.[9] This hands-on research does not mean leaving out historical analysis. The experimentations used here are able to show ways in which history is active in contemporary digital structures. Images, datasets, algorithms and computer vision in general are the continuation of historical imperialist practices and some experimentations can highlight that. Finally, I

2 Anthony McCosker and Roman Wilken, *Automating Vision: The Social Impact of the New Camera Consciousness*, New York: Routledge, 2020, p. 3.

3 Vinay Uday Prabhu and Abeba Birhane, 'Large image datasets: a pyrrhic win for computer vision?', in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 1536–46.

4 Li Fei-Fei, Jia Deng, Olga Russakovsky, Alex Berg and Kai Li, 'ImageNet', at https://image-net.org (accessed 24 March 2022); Mark Everingham, Luc van Gool, Chris Williams, John Winn and Andrew Zisserman, 'Pascal-Voc', at http://host.robots.ox.ac.uk/pascal/VOC (accessed 24 March 2022).

5 Questionnaire applied from 1 July to 1 August 2021.

6 Hazel Smith and R.T. Dean, *Practice-Led Research, Research-Led Practice in the Creative Arts*, Edinburgh: Edinburgh University Press, 2009, p. 5.

7 Smith and Dean, op. cit. (6), p. 6.

8 Annette N. Markham and Gabriel Pereira, 'Analyzing public interventions through the lens of experimentalism: the case of the Museum of Random Memory', *Digital Creativity* (2019) 30(4), pp. 235–56.

9 Markham and Pereira, op. cit. (8), p. 236.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

should point out that rather than individual actions, these experimentations were carried out alongside programmers from the Group on Artificial Intelligence and Art (GAIA–C4AI, Inova USP).[10]

To emphasize the relationship between historical perspective and experimentalism, the sections of this article are determined from the relationship between historical colonial practices and specific experimentations that evidence the continuity of this past – now in the guise of algorithms. In the first section, I look exclusively at the visual content of these images in a hand-curated way, reviewing the thousands of folders of LSVDs. The lack of diversity in LSVDs is highlighted through Experimentation 1, which overlays images from Flickr-Faces-HQ. Then I focus on LSVD images that relate to nature. Experimentation 2 uses a programming code to show the colonial extractive logic behind computer vision in the face of these natural scenarios. For the next section, I created an interactive website that turns our attention from images to the ways they are organized by crowdworkers (Experimentation 3 – Exch w/ Turkers). People like Pedro, Sonia and Anand (some of the 700,000 microworkers on Amazon Mechanical Turk) must be considered in a critical investigation of computer vision. Focusing on the human labour behind organizing images will show how this precarious work contributes to the historical perpetuation of colonial, gender and racial norms. Responding to this, the next secion discusses an alternative crowdsourcing project. The *demonumenta* project (Experimentation 4) tested new ways of tagging images to shift these historical perpetuations of power through images using historical paintings. The collectively constructed machine-learning training dataset drew on two previous non-hegemonic initiatives using images in 1960s Latin America: Paulo Freire's literacy method and the Mail art exchange coordinated by art critic Walter Zanini. His Mail art exchange served as an important reference for the fifth and last experimentation, with results that appear (explicitly and implicitly) throughout the article and that materialize images from AI datasets in postcard format. These were sent by mail, in threes, to eighty-five researchers, artists and crowdworkers during the second semester of 2021. Feedback from this group was fundamental to establishing new ways of understanding the images.

As the conclusion will show, including images that train computer vision in the debate around machine learning allows me to discuss another layer of the historical practice of categorizing and levelling the visual world. But it can also contribute to the larger debate on a contemporary notion of rationality that is increasingly 'data-centric'.[11] For James Bridle, we are in the 'New Dark Age', where 'computation does not merely augment, frame, and shape culture; by operating beneath our everyday, casual awareness of it, it actually *becomes* culture'.[12] Here, I hope to better understand this stage of Western scientific positivity through some of its images and related practices, including the routines of microworkers.

Viewing images: white hands that fish

Realizing that LSVD images are levelled for organizational purposes is an important first step towards understanding them as a perpetuation of historical practices and normativities. In the 1970s television series *Ways of Seeing*, art critic John Berger argues that the process of seeing 'is less spontaneous and natural than we tend to believe', and further limited by the fact our eyes can only 'be in one place at a time'. This has changed with

10 The Group on Artificial Intelligence and Art (GAIA) is a network of researchers and artists interested in reflecting on and debating contemporary digital infrastructure in an experimental and artistic way.

11 Paola Ricuarte, 'Data epistemologies, the coloniality of power, and resistance', *Television & New Media* (2019) 20, pp. 350–65.

12 James Bridle, *New Dark Age: Technology and the End of the Future*, New York: Verso Books, 2018, p. 39, original emphasis.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

advancing technology, including widespread use of the photographic camera; now images can 'travel across the world'. Berger died in 2017 and never wrote about images that train AI, but the warning he gives at the end of the first episode is still relevant. Looking straight into the camera, he says, 'You receive images and meanings which are *arranged* ... be sceptical of it.'[13] Relating Berger's warnings to computer vision, Mitra Azar, Geoff Cox and Leonardo Impett consider them a prescient reaction to an increasingly non-human contemporary visual field marked by 'alienated forms of social interaction'.[14]

The empirical work of looking (sceptically) at a subset of the millions of images in LSVDs allows us to confront an idea that some of their creators defend: that images, when compiled, are attempts to understand the world. In fact, computer vision does not do this, even if some experts view its datasets as 'progress towards genuine scene understanding', as a 'new state of the art in class-conditional image synthesis' and 'large-scale groupings'.[15] They might be large-scale, but a close look at the images present in the many folders of four well-known LSVDs – ImageNet, Flickr-Faces-HQ, Tiny Images and Google Open Image – shows that the promise LSVDs offer to understand the world is not in fact fulfilled.

This idea of a limited understanding of the world can be seen in one of the best-known face image sets for training computers. The Flickr-Faces-HQ dataset contains about 70,000 images of human faces and is widely used to train algorithms capable of creating deep fakes. Algorithmically generated faces are part of the advancements in computer vision since 2014, with the development of a machine-learning infrastructure called Generative Adversarial Network – GAN – in which two networks compete (the generator and the discriminator).[16] More recently, this has enabled the production of more realistic images from StyleGAN and its successor StyleGAN2.[17]

This Person Does Not Exist (TPDNE) is one of the models that results from using GANs. When GAIA researcher Lucas Nunes and I learned of it, we decided to try something simple: for one week, we ran this model endlessly to analyse what kind of faces came up as results. We generated thousands of faces and decided to look at them very quickly in sequence, projecting them onto the wall. The effect was a little nauseating, but it helped us see they were pretty much all white people's faces. After that, we carried out a set of clustering processes (Experimentation 1) that help exemplify limits of the models created from the Flickr-Faces-HQ dataset.[18] Nunes created cluster images by overlapping *N* random and unrepeated fake human faces available in a set of 4,500 images in TPDNE.

The results are near-identical faces (Figure 1) representing a person with seemingly white skin, dark eyes and brown hair, predominant characteristics regardless of how

13 John Berger, *Ways of Seeing*, Episode 1, London: BBC, 1972.

14 Mitra Azar, Geoff Cox and Leonardo Impett, 'Introduction: ways of machine seeing', *AI & Society* (2021) 36, pp. 1093–1104, 1093.

15 'Open Image Dataset', at https://storage.googleapis.com/openimages/web/index.html (accessed 24 March 2022); Andrew Brock, Jeff Donahue and Karen Simonyan Brock, 'Large scale GAN training for high fidelity natural image synthesis', https://arxiv.org/abs/1809.11096 (accessed 24 March 2022); Li *et al.*, op. cit. (4).

16 Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, 'Generative adversarial nets', *Advances in Neural Information Processing Systems* (2014) 27, pp. 2672–80.

17 Tero Karras, Samuli Laine and Timo Aila, 'A style-based generator architecture for generative adversarial networks', *EEE Transactions on Pattern Analysis & Machine Intelligence* (2021) 12, pp. 4217–28; Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen and Timo Aila, 'Analyzing and improving the image quality of StyleGAN', *Conference on Computer Vision and Pattern Recognition* (2020), pp. 8110–19.

18 Lucas Nunes, Bruno Moreschi and Amanda Jurno, 'Which faces can AI generate? Normativity, whiteness and lack of diversity in This Person Does Not Exist', *Beyond Fairness: Towards a Just, Equitable, and Accountable Computer Vision Conference* (2022), pp. 1–9.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

Figure 1. Experimentation 1. A set of three images resulting from the overlapping of different images from the TPDNE. Clusters A, B and C were generated from one hundred, five hundred and one thousand images respectively, taken from the Flickr-Faces-HQ dataset at random and without repetition. Credits: Lucas Nunes/GAIA–C4AI, InovaUSP.

many fake faces are overlapped. Experimentation 1 shows that artist Giselle Beiguelman was correct in arguing that deep fakes are a kind of 'eugenics of the gaze', proving that these generated images are not harmless.[19] In addition, LSVDs like Flickr-Faces-HQ are created without the consent of those whose faces are used to train these algorithms. In the Big Data era, advancements in privacy resulting from international efforts in reaction to Nazism (such as the 1947 Nuremberg Code and the 1964 Declaration of Helsinki) 'have gradually been eroded'.[20]

As Experimentation 1 shows, models such TPDNE may contain a lot of data, but not necessarily much diversity; they are not actually suitable for providing a broader understanding of the world.[21] The main reason is that LSVDs like Flickr-Faces-HQ are characterized by 'the absence of critical engagement with canonical datasets' – women, as well as racial and ethnic minorities, are negatively impacted.[22] Recent studies have shown how algorithms based on machine learning discriminate against people on a phenotypic basis and offer readings of the world based on normative standards, almost always related to consumption and hegemonic cultures.[23] What does not conform to this strict logic is the unexpected, a type of data that is not welcome in commercial AI.

Experimentation 1 shows the importance of carefully analysing computer vision images, as done with TPDNE. It is a simple approach, but one important to understanding the way machines see, enabled in part by these LSVD images. It is not a way of apprehending, but of prioritizing and decontextualizing, specific elements in visual content through the logic of categorization. This categorization logic can be better understood by focusing

19 Giselle Beiguelman, *Políticas da Imagem: Vigilância e Resistência na Dadosfera*, São Paulo: Ubu Editora, 2021, my translation.

20 Prabhu and Birhane, op. cit. (3), p. 1.

21 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: NYU Press, 2018; Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York: St Martin's Press, 2017; Cathy O'Neil, *Weapons of Math Destruction*, Largo: Crown Books, 2016; Catherine D'Ignazio and Lauren F. Klein, *Data Feminism*, Cambridge, MA: MIT Press, 2020.

22 Prabhu and Birhane, op. cit. (3), p. 2.

23 Joy Buolamwini and Timnit Gebru, 'Gender shades: intersectional accuracy disparities in commercial gender classification', *Conference on Fairness, Accountability and Transparency* (2018), pp. 77–91; Kate Crawford and Trevor Paglen, 'Excavating AI' (2019), at https://www.excavating.ai (accessed 24 March 2022); André Mintz and Tarcizio Silva, 'Interrogating vision APIs', *Smart Data Sprint: Beyond Visible Engagement* (2019), pp. 25–54.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

on LSVD images of nature. In the ImageNet dataset, a significant part of the 'train' subfolder in the 'imagenet12' folder was created to separate elements of fauna and flora based on visual characteristics. Snakes are separated into seventeen folders, roses into eight, elephants into two, and so on. Each ImageNet folder has exactly 1,252 images, with rare exceptions.

The practice of specifying becomes even more evident when we take into account that there are many computer vision datasets constituted for species and breeds in websites such as the UCI Machine Learning Repository.[24] The Stanford Dogs Dataset, for example, is available for download with 20,580 images, divided into 120 dog breeds.[25] Although it may be delightful to get lost in images of different dogs, by compiling them in this way this system implies an idea of totality that is an oversimplification; after all, where do stray dogs fit into this dataset?

Michel Foucault explains that the practice of taxonomy is not about discovering the names of things, but about the world only containing things with names. In Foucault's words, taxonomy implies 'a certain continuum of things'.[26] Kate Crawford also reminds us that this classificatory continuum of taxonomy is not just a movement in itself, but an instrument of power.[27] In this instrumentalization, identified as 'systems of classification' by Vinay Prabhu and Ababa Birhane, power operates in an asymmetrical way, so that what is normal or acceptable is 'often dictated by dominant ideologies'.[28]

The division logic of LSVDs such as ImageNet and Tiny Images is related more directly to the fact that its categories were taken from a lexical database of semantic relations between words: WordNet, a kind of dictionary of categories. WordNet is but one of the many possible formalizations of this broader logic process. In Google Open Image, for example, there are 1,900,000 images divided into categories, but no information about their origin. In the Leaf Dataset, forty different plant species were photographed over a contrasting background, causing them to look like images from the Rorschach test, a letdown for nature lovers. But for machines, there are specific attributes for each leaf image: 'class, specimen number, eccentricity, aspect ratio, elongation, solidity, stochastic convexity and smoothness'.[29] These datasets, and so many others associated with computer vision, are examples of the applicability of what the botanist and zoologist Carl Linnaeus popularized in the eighteenth century: the binomial nomenclature created by Gaspard Bauhin and scientific classification. Linnaeus's taxonomic logic is embedded in machine learning: from specific to general and from general to specific. Furthermore, the fact that important computer vision datasets 'inherit' the categories of previous structures – like WordNet or more general modes of taxonomic practice – shows the importance of understanding them beyond their individual structures: more than systems, we are discussing systemic continuities.[30]

24 Dheeru Dua and Casey Graff, 'UCI Machine Learning Repository', at https://archive.ics.uci.edu/ml/index.php (accessed 24 March 2022).

25 Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao and Li Fei-Fei, 'Stanford Dogs Dataset', at http://vision.stanford.edu/aditya86/ImageNetDogs/main.html (accessed 24 March 2022).

26 Michel Foucault, *The Order of Things: An Archaeology of the Human Science*, 2nd edn, London and New York: Routledge, 2005, p. 80.

27 Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven, CT: Yale University Press, 2021.

28 Prabhu and Birhane, op. cit. (3), p. 6; Susan Leigh Star and Geoffrey C. Bowker, 'Enacting silence: residual categories as a challenge for ethics, information systems, and communication', *Ethics and Information Technology* (2007) 9, pp. 273–80.

29 Pedro Silva, André Marçal and Rubim Almeida da Silva, 'Evaluation of features for leaf discrimination', in Kamel M. and Campilho A. (ed.), *Image Analysis and Recognition*, Berlin: ICIAR, Springer, 2013, pp 197–204.

30 Prabhu and Birhane, op. cit. (3), p. 3.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

Interpreting this process of specification by images in terms of Linnaeus's modern taxonomy places AI in a historical perspective, beyond the confines of engineers and programmers. Authors such as Aníbal Quijano are fundamental for deeper understanding of the relationship between coloniality and modernity. According Quijano, with the conquering of societies in places like Latin America and the African continent, there begins a new world characterized by a 'global power covering the whole planet'.[31] This violent process is often described as 'objective' and 'scientific', something that decolonial studies vehemently refute, as it strips away its historical layers. Quijano calls for us not to consider this process a natural phenomenon, but a historical process of power perpetuation. These considerations help us to understand the transformation that Matteo Pasquinelli describes in his Nooscope Manifesto: the project of mechanization of human reason that characterized previous centuries is now a corporate regime of knowledge extractivism and epistemic colonialism.[32] My argument here serves to place computer vision and its ways of classifying the world in this same historical perspective.

From that perspective, LSVD images are imbricated in the many colonial projects under way between the sixteenth century and the late nineteenth, with their concomitant statistical processes, censuses and division of colonial territories, such as the Treaty of Tordesillas (1494) or the Berlin Conference (1884–5), where the South American and African continents were divided, respectively, from a distance.[33] Imperialist initiatives such as these are arguably the historical antecedent to the decontextualized practices of computer vision. It is not hard to relate the straight geographical lines resulting from these two colonial episodes to the rectangles in computer vision referred to by experts as 'meaning boxes' – the lines of the latter also cutting out pieces of territories.[34]

What I propose in Experimentation 2 is to consider that this view that selects and extracts from a distance remains active in the practice of 'seeing' in AI. The historical continuum identified by Foucault and updated in considering digital infrastructures by Crawford is perceptible, for example, in the Python programming code created by GAIA programmer Bernardo Fontes and myself. This code was used to read the data present in the 'meaning boxes' of nature images in the Google Open Image dataset. The code was written to identify these selections and discard them, resulting in images that are no longer important for computer vision, as they were now made just by their unselected areas (Figure 2).

The incomplete images provided by Experimentation 2 expose the practices that LSVDs impose on the natural world. These informational gaps not only indicate the idea of extractivism and exploitation, but also provoke us to think about the systemic destructuring that computer vision performs in territories, animals, plants and so on. We call this reverse-engineering experimentation 'decanonization', as technical articles about computer vision commonly use the term 'canonization' to describe the process of the machine seeing and choosing what should be highlighted in each image.[35]

Evidencing the white areas in these images also encourages me to think about who operates in these extraction spaces. As I write this article, I see an overwhelming majority

31 Aníbal Quijano, 'Coloniality and modernity/rationality', *Cultural Studies* (2007) 21, pp. 168–78.

32 Matteo Pasquinelli and Vladan Joler, 'The nooscope manifested: AI as instrument of knowledge extractivism', *AI & Society Journal* (2021) 36, pp. 1263–80.

33 Arjun Appadurai, *Fear of Small Numbers: An Essay on the Geography of Anger*, Durham, NC: Duke University Press, 2006.

34 Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein and Li Fei-Fei, 'Visual Genome: connecting language and vision using crowdsourced dense image annotations', *International Journal of Computer Vision* (2017) 123, pp. 32–73.

35 Krishna *et al.*, op. cit. (34), p. 32.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

Figure 2. The decanonization process in two images of the Open Image dataset. The white rectangle corresponds to the area of the image that Google's AI considers relevant. Credits: Bernardo Fontes and Bruno Moreschi, GAIA–C4AI, InovaUSP.

of white men in the LSVDs. Little by little, I realize I am empirically confirming the fact that most of the images that train AI are taken from Flickr – a social network consisting of images of contemporary life and social practices, mostly from the United States, which

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

Figure 3. One of 1,252 images of white men holding their freshly caught fish in an Imagenet folder. The image is low-resolution (4KB), like many others in this dataset. Credits: https://image-net.org.

produces 30 per cent of Flickr content.[36] In this tour of normative US images (Halloween pumpkins, people eating pizza, road signs in English), I fix my eyes on a folder that seems like a synthesis of the idea that these images serve as material to enable domination.

ImageNet subfolder 0001 (imagenet10 > train) contains 1,252 images of people the moment they catch a fish. In those pictures, there are always smiling white men showing off their catch like trophies (Figure 3). There are no fishermen folders in Visual Genome, Tiny Image and Open Image, but they are often present in a more dispersed way. Added to this set of fisherman images are an infinite number of images in other folders of these four LSVDs that show small animals, vegetables and minerals, held in the palms of white hands. These images speak louder than words – the many fish caught by these hands are also possible representations of the historical domination of white men from the northern hemisphere over nature.

Looking at the images that make up LSVDs can give us an important understanding of their logic. A visual analysis (e.g. audit cards and institutional review boards) of their content can interfere in the curation process of what should and what should not be part of these datasets.[37] These actions are not just welcome, but urgent. More careful curation of images and their categories is part of a review that could result in a less problematic computer vision field. However, as we intend to show below, there are, in addition to the millions of images in computer vision datasets, a set of human practices for organizing, labelling and validating those images that should also be looked at carefully.

Seeing through images: sleeping and tagging in the same room

Experimentation 3, called Exch w/ Turkers, is a website that allowed the general public to chat with Amazon Mechanical Turk microworkers (or turkers) for twenty-two days, to raise awareness around the fact that although often presented as autonomous, AI 'is made of people'.[38] Visitors communicate with turkers through website chats to

36 'Similar Web', at www.similarweb.com/website/flickr.com/#overview (accessed 24 March 2022).

37 Prabhu and Birhane, op. cit. (3), pp. 1, 10.

38 Lilly Irani, 'The labor that makes AI magic', White House/NYU AI Now Summit Talk, 7 July 2016, at www.youtube.com/watch?v=5vXqpc2jCKs (accessed 24 March 2022); Gilles Bastin and Paola Tubaro, 'Le moment big

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

understand the daily routine of those behind HITs – used, among other things, to organize and tag images for LSVDs.[39] The work of turkers is characterized by low remuneration and anonymity in the face of what is being built.[40]

Once the interactive part of the Exch w/ Turkers was complete (its static content remains online), I continued a weekly exchange with the turkers where we went back and looked at their responses in the chats to further deepen their considerations. They had access to a file with the chat transcripts where they could insert footnotes for anything that they might like to complement or reconsider.[41]

In our exchange, I insisted that they help me in the process of understanding the images they tag and describe. It is relevant to note that when talking about the images, the turkers almost never described them, but focused more on how they were organized. One of the participants, Pedro, mentioned more than once that he relates the HITs involving images to the sound of two plug-ins (HIT Catcher and HIT Finder) that he installed on his browser: 'I like these HITs because they are quick. But, at the same time, it's common for this sound to wake me up at night. If it's a simple image-tagging task, for example, I'll get up to do it'.[42]

Another turker, Sonia, works on the platform to supplement household income and buy imported toys for her two children, like Baby Alive dolls and Lego boxes. To reconcile her work as a turker with being a housewife, she set a computer up in the living room, bedroom and kitchen. They are always on, but, for HITs with images, Sonia prefers to use the big monitor in the living room: 'Not just because I see the images in a larger size, but because it is also a way to get the kids involved. Sometimes tagging images feels like a game to them'.[43] We can relate this comment about working as entertainment to the idea of gamification that characterizes online work platforms.[44]

A third turker, Anand, was more succinct. He described the images used in his work as trivial and something he forgets about once he finishes a task. He did not send me screenshots of his monitor as he worked, but sent me a very valuable image for our analysis: a photograph of the place that is both his bedroom and his office (Figure 4). Months later, he came back to it in a conversation, asking if this scenario could be different.[45] The question Anand posed is an attempt to contemplate fairer and more efficient ways to train machines. I argue that this image needs to be looked at as critically as the images found in LSVDs. This is because the images provided by turkers show precisely what LSVDs hide: the precarious work carried out by real people.

---

data des sciences sociales', *Revue française de sociologie* (2018) 59, pp. 375–94. Far from being an inexpressive segment of the digital economy, microwork like that carried out by Amazon Mechanical Turk (AMT) is growing and should be understood as a consequence of a growing process of the datafication of society. AMT is the best-known online working platform for this new algorithmic achievement of capitalism; however, it is not the only one. Clickworker, Figure Eight, Fiverr and JobBoy are some of its competitors.

39 'Exch w/ Turkers', at https://exchanges.withturkers.net (accessed 24 March 2022), carried out during my research in the Histories of AI: A Genealogy of Power (University of Cambridge/Mellon Sawyer Seminar).

40 Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*, Boston, MA: Houghton Mifflin Harcourt, 2019; Kinda El Maarry, Kristy Milland and Wolf-Tilo Balke, 'A fair share of the work: the evolving ecosystem of crowd workers', *Proceedings of the 10th ACM Conference on Web Science* (2018), pp. 145–52.

41 As with the first part of the project (website), the crowdworkers were paid – sixteen dollars per hour, an amount decided on collectively.

42 Personal conversation, 12 March 2021.

43 Personal conversation, 9 October 2021.

44 Jamie Woodcock and Mark R. Johnson, 'Gamification: what it is, and how to fight it', *Sociological Review* (2017) 66(3), pp. 542–58.

45 Personal conversation, 20 November 2021.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

Figure 4. Turker Anand's bedroom and workspace in New Delhi, India. Credits: Exch w/ Turkers.

Organized by microworkers like Anand, Pedro and Sonia, Visual Genome is a computer vision model that promises to be many things: 'an ongoing effort to connect structured image concepts to language', 'a state of art', 'a knowledge base', 'an unprecedented innovation in the field of computer vision', and so on.[46] According to its creator, Visual Genome is a milestone in computer vision because it solves tasks that computers have previously performed poorly; that is, 'cognitive tasks'.[47] Cognition, for these new image specialists, is a term for automated processes capable of describing images based on straightforward questions and answers. While this is definitely not cognition, it is no small feat. With Visual Genome, it is possible to identify not only that there is a person and a vehicle in an image, but also that there is an elderly man riding a horse-drawn carriage. For this, 108,000 images were marked so that each one had an average of thirty-five objects, twenty-six attributes and twenty-one listed pair relationships.

The article that explains Visual Genome has seventy-two paragraphs, thirty-seven figures (twenty in charts), nine tables with numbers and a mathematical formula, but only two short paragraphs relate not to the image annotations but to those who write them, the turkers. This is superficially explained in a topic called 'Crowdsourcing strategies'.[48] There, it is revealed that Visual Genome images were tagged entirely by turkers. In total, 33,000 workers from Amazon Mechanical Turk (93.02 per cent from the US and 6.98 per cent from other countries, divided equally between men and women) were responsible for the existence of this computer vision model, having carried out 800,000 HITs. For this, they received between six and eight dollars an hour. The result of this collective work is 108,077 images with 5.4 million regions described, 1.7 million answers and associated questions, 3.8 million identified objects and 2.3 million relations between these objects.

In considering Anand's photograph of his room/workspace and highlighting the low visibility of his work in the technical articles on Visual Genome, I aimed to show what is not commonly seen in LSVDs. Shedding light on the invisibility of turkers invites us

46 Krishna *et al.*, op. cit. (34).
47 Krishna *et al.*, op. cit. (34).
48 Krishna *et al.*, op. cit. (34), p. 43.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

to consider that part of the problem with computer vision is not only the limitations of the visual content, but the way in which it is operated/manipulated to be used as training data. Below, encouraged by Anand's question – can this be different? – I intend to think of an alternative way to train computer vision.

Revealing 'coloniality of power' with images from a historical museum

In highlighting the problems brought on by training computer vision in a decontextualized and precarious way, and rethinking alternatives, I have been particularly inspired by the expansion of contemporary art from the 1960s, especially the 'dematerialization of art'.[49] Attention to this history can inform contemporary debate about AI through the non-hegemonic collaborative mediation practices that it made possible. The exploration of new social communication technologies during the last four decades of the twentieth century allowed for artistic practices – partly using Futurism and Dadaism as references – to go beyond the notion of artwork as an object and thereby explore how art infiltrates everyday life. This change opens a field of opportunities for art to act as a cognitive instrument: epistemology replaces the abstract formalism of modern art.[50] This complex ontological shift deeply interested the conceptual art of the time and was described by art critic and curator Walter Zanini as the 'crisis of the art object and its substrate'.[51]

Some artists and educators from that period were trying to address Norbert Wiener's concern that understanding society is only possible through the investigation of communication.[52] In some countries, referred to then as 'Third World', the transformation of art was more than a self-directed institutional critique, as was partially the case in the US and Europe.[53] In Latin American countries such as Argentina and Brazil (from where I am writing), the dematerialization of art helped several artistic and educational projects open up space for collaborative exchanges as a reaction to political and social challenges. Some of them were only possible because of non-traditional ways of dealing with images. They interest me as possible alternative methodologies for machine-learning datasets.

Two examples might explain this further. The first is in the field of education, where the adult literacy teaching methodology proposed by Brazilian educator Paulo Freire in the 1960s made use of 'generative words' (*palavras geradoras*) – words related to the everyday lives of adult students, most of whom were lower-class rural and urban workers. The Freirian method proposes creating posters with vocabulary selected in collaboration with students, related to their everyday lives. In this learning context, images are used not as decontextualized illustrations, as is usually the case in textbooks (and in LSVDs), but as visual syntheses of collective and local mediations.

Freire initially tested his method with five rural workers, three of whom, according to him, learned to read and write in little more than a day. This feat caught the attention of the team under Brazilian president João Goulart, resulting in a national literacy campaign that benefited 2 million people from 20,000 different reading circles. However, the 1964 civil–military coup interrupted the project and Freire was forced to go into exile in Chile. The contextualized, localized and collaborative approach that Freire proposed

49 Frank Popper, *From Technological to Virtual Art*, Cambridge, MA: MIT Press, 2006.

50 Michael Newman and Jon Bird, *Rewriting Conceptual Art*, London: Reaktion Books, 1999.

51 Walter Zanini, *Vanguardas, Desmaterialização, Tecnologias na Arte*, São Paulo: WMF Martins Fontes, 2018, p. 305, my translation.

52 Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society*, New York: Da Capo Press, 1988.

53 Cristina Freire, *Arte Conceitual*, Rio de Janeiro: Editora Zahar, 2006, pp. 10–33; Catherine Spencer, 'Navigating internationalism from Buenos Aires: The Centro de Arte y Comunicación', *ARTMargins* (2021) 10, pp. 50–72.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

informs my speculation about new ways of tagging images and, consequently, constituting less problematic datasets.

The second example of an alternative use of images, also in 1960s Latin America, was the Mail art exchange between artists in countries under dictatorships, such as Argentina, Brazil, Chile, Colombia, El Salvador, Mexico, Uruguay and Venezuela. Postcard art (drawings, photographs, instructions for performances) was made by art students in Brazil, and the exchange was coordinated by Walter Zanini, a professor at the University of São Paulo and curator at the Museum of Contemporary Art USP. The art critic Cristina Freire described this international 'solidarity network' as a social reaction in image form: 'If collective memory was threatened in public spaces, it was reinforced in alternative networks through virtual encounters'.[54] Zanini considered this exchange project via artistic postcards part of 'an appeal to immaterial reality' based on the means of communication of the time.[55]

Some sixty years later, at the same Brazilian university where Zanini and his students helped create the Mail art exchange in Latin America, now as a distance-learning experience during the COVID-19 pandemic, the *demonumenta* project (Experimentation 4) brought together around thitry undergraduates, graduates and professors from the School of Architecture and Urbanism at the University of São Paulo to empirically think of a new way to create machine-learning training datasets.[56] The images used were history paintings from Museu Paulista, an institution inaugurated on 7 September 1895 (Brazilian Independence Day) that was and is fundamental for the construction and perpetuation of the history of Brazil narrated by the São Paulo elite.

For the *demonumenta* dataset, we chose fifty tags to categorize the museum works, taking into account Freire's method of considering our specific context. General categories were defined, such as 'sky', 'fauna' and 'flora', as well as more specific ones. Categories took into account a decolonial perspective such as 'white man',' 'indigenous man', 'black man', 'white woman', 'indigenous woman', 'black woman', 'indigenous child', 'black child', 'white child', 'enslaved', '*bandeirante*', 'military', 'coffeemaker', 'farmer' and so on. These categories resulted in our tagging process, which involved a repetitive overlapping of layers that revealed historical correlations (Figure 5). The category 'white man', for example, is always associated with other categories like 'coffee producer', 'politician', 'military', and/or '*bandeirante*'.[57] From these categories it is also possible to see that while white men tend to be represented in portrait paintings (their names included in the titles), indigenous and black people almost always appear in groups, occupying smaller areas of the paintings.

Working with the codes associated with these files helps reveal how images are selected and arranged differently in LSVDs than those activated by Freire and Zanini. In fuelling computer vision, LSVD images act as part of a 'culture' that datasets like ImageNet have helped create in the AI community. Part of this culture is to reduce complex interpretations of images, their elements and their interactions. Thus proposing exchanges and other not-so-traditional approaches is a way to reveal that this culture of levelling categories 'simplifies and freezes nuanced and complex narratives, obscuring political and moral reasoning behind a category'.[58]

54 Bruno Sayão, 'Solidariedade em Rede: Arte Postal na América Latina', master's thesis in aesthetics and art history, Universidade de São Paulo, 2016, my translation; Freire, op. cit. (53), p. 69.

55 Zanini, op. cit. (51), p. 109, my translation.

56 'demonumenta', at http://demonumenta.fau.usp.br (accessed 12 November 2021).

57 Portuguese descendants who, from the beginning of the sixteenth century, advanced into the interior of South America in search of gold and silver, engaging in the genocide of indigenous communities. References to the *bandeirantes* not only are found in museum paintings; they are also the names of many parks, schools, streets and highways in Brazil.

58 Prabhu and Birhane, op. cit. (3), p. 2.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

Figure 5. An example of overlapping tagging that indicates historical associations – here, the category 'white man' is connected to that of 'politician'. Portrait of Dom Pedro I, 1902, by Benedito Calixto. Credit: José Rosael/Hélio Nobre/Museu Paulista USP.

Unlike LSVDs, we explicitly indicated the origin of the choices made in constructing the *demonumenta* dataset. We were aware of our own limitations and the world views we brought to the classification process. The '*quilombo*' category is a good illustration of this more critical tagging process. *Quilombo* is the name given to communities formed

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

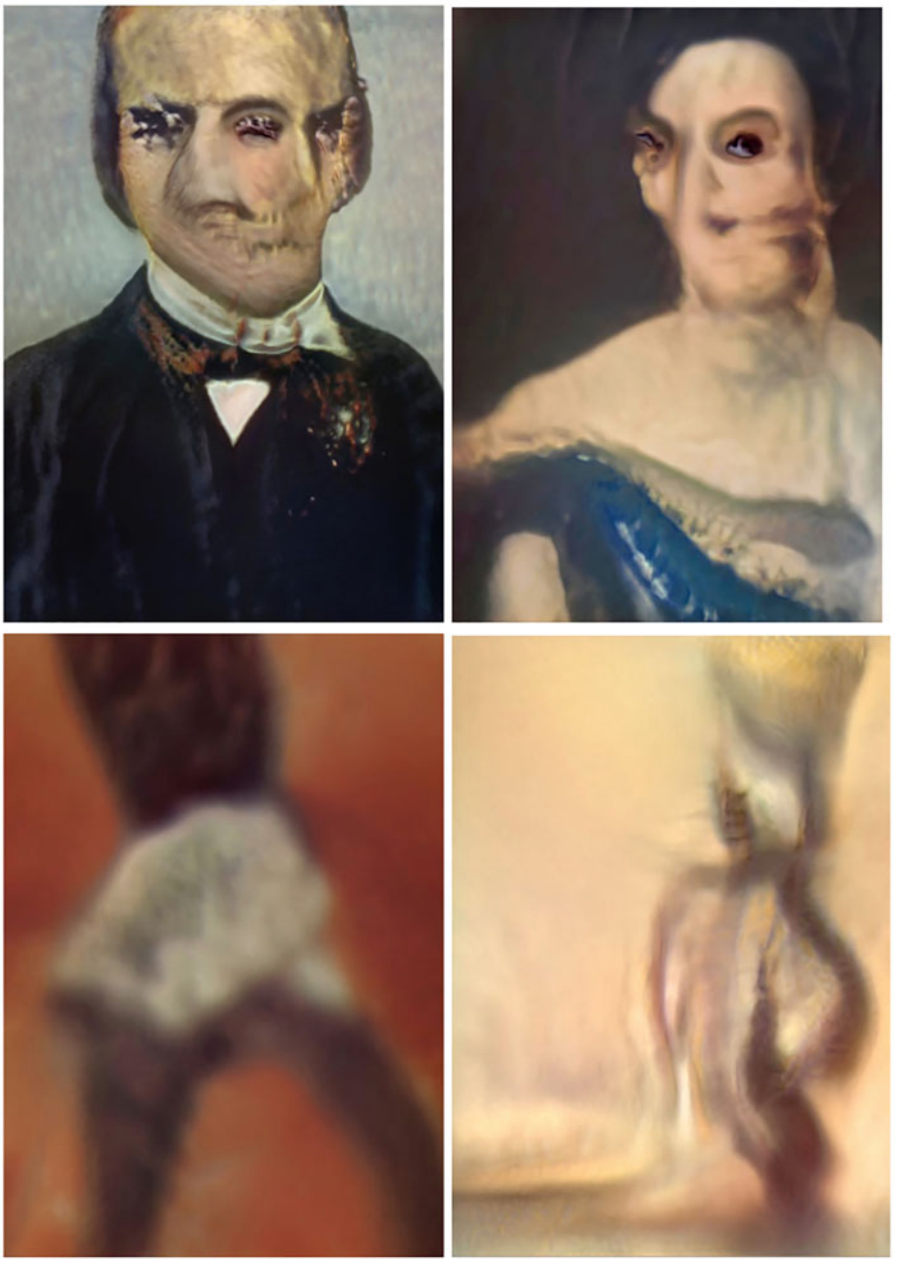

Figure 6. Results of the experiences with GANs in the project dataset. From left to right: a white man and white woman, a black man and black woman. This stage of the project was carried out in partnership with Giselle Beiguelman and Bernardo Fontes, with support from the Intelligent Museum artist residency, Center for Art and Media/ZKM Karlsruhe.

by fugitive enslaved people in Brazil. Initially, we created it because it was an element of black resistance. However, precisely for this reason, we did not find any representation of *quilombos* in the collection of the Museu Paulista – an institution in which practically all directors have been white, as well as most of the artists in its collection. Even with no

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

corresponding images in the category *quilombo*, the *demonumenta* project team considered it important to maintain it to signal a possibility for future contemplation in other contexts. Even if this tag is never filled in other data collections, this absence will never be merely an empty space in our dataset; quite the contrary: it is revealing of how whiteness impacts the decision of what is and is not relevant in images.

The *demonumenta* dataset served as more than just an alternative tagging process. It also enabled results with GANs, which allowed us to use AI to identify what were the most characteristic elements of the categories perpetuated in the museum's collection and then to create composite images of them. The algorithmic images of white man and white woman produced by GANs via the *demonumenta* dataset are represented in formal clothes (including deep-fake ties and dresses). Comparatively, the algorithmic images of black man and black woman do not have a face and are in a working stance (the man is pulling something and the sensualized woman is carrying a vase on her head – Figure 6). Such results shift the discussion of artistic representation from that of an individual artist's decision to a broader aesthetic structure – part of 'a certain continuum of things', in Foucault's words (see above). The experience also evidences what Quijano refers to as the 'coloniality of power'; in other words, that the relations and discourses imposed during the colonial period continue beyond colonization itself, refined through new modalities.[59]

It is also possible to relate the algorithmic images of black man and black woman resulting from the *demonumenta* project (Experimentation 4) with the workers participating in the Exch w/ Turker project (Experimentation 3). Despite being different (one is slavery per se, the other a precariousness of work), the labour exploitation in each has in common the anonymization of the workers. In the forced labour that so characterized colonial Brazil and in the current way of training machines, faces, names and personal life stories of workers are disregarded.

By carrying out a computer vision process from start to finish – in other words, selecting images, tagging some of their parts and training algorithmic models with this material – *demonumenta* shows that treating training datasets as collective and contextualized experiences can be important for exposing and interrogating the colonial and other logics of power informing the datasets on which computer vision is trained. Even though AI and its datasets present themselves to the world almost as 'magic tricks', they are the result of the sedimentation of the historical discourses that their data carry.[60] This junction between colonization and the official cultural apparatus (including official images) is the real lexical and imagery basis for the categories present in computer vision.

Lastly, the notion that datasets are built collectively expands computer vision beyond its visual surfaces. Acknowledging those who organize the images encourages us to expand the idea of only looking at images critically to actions that also transform the field of computer vision, including its practices, spaces and working conditions. In the end, we are not only talking about images; we are talking about people and their relationship with images.

Conclusion

The experimentations described here help to highlight three important points (that are not so evident at first glance) about the subfield of AI concerned with computer vision: (1) how LSVD images reproduce logics of extractivism from colonial projects and modern taxonomy to train computational models capable of acting as instruments of power to

59 Aníbal Quijano, *Ensayos en Torno a la Colonialidad del Poder*, Buenos Aires: Ediciones del Signo, 2019.
60 Ed Finn, *What Algorithms Want: Imagination in the Age of Computing*, Cambridge, MA: MIT Press, 2017.

https://doi.org/10.1017/bjt.2023.6 Published online by Cambridge University Press

maintain historically dominant ideologies; (2) the fact that these images only become training data for this perpetuation of power after being reorganized in a decontextualized way by thousands of microworkers, updating historical forms of exploiting human labour; and (3) the possibility of using these images and official images (such as those in historical museums) to become datasets for machine learning in less problematic ways through collaborative, educational and artistic projects inspired by previous counterhegemonic initiatives in Latin America.

In the fifth and last Experimentation, which involved sending postcards with LSVD images by mail to a select group of experts and microworkers in computer vision, much of the feedback supported the idea that we are dealing with images that not only enable computers to 'see', but also perpetuate ways of understanding the world through historically active hegemonic norms. These LSVD images are part of a larger mechanism of power that was operating long before the current AI trend.

When visual-culture professor Iara Schiavinatto received her three postcards, she looked at them for about three weeks as they sat next to her keyboard. In a videoconference giving feedback, I was intrigued that she used the word 'violence' many times to describe them. The images sent to her looked so innocuous: two illustrations of plants, and a man playing a harmonica. Schiavinatto explained that this did not mean that these images depict violence per se, but that they are part of a 'management of violence' and 'this is what our eyes need to pay attention to'. Although filled with cute kittens, the images gathered in LSVDs have 'a brutal disciplinary effect'.[61] While Schiavinatto and I were talking, the turker Pedro sent me another screenshot of his work. It was a task on Amazon Mechanical Turk with many smiling faces, in which he had to answer how happy the faces were – on a scale of 0 to 10.

As Pedro rated the smiles, he himself was unsatisfied with his working conditions and the amount he was paid for each HIT (ten cents). When creating these datasets, our own smile and satisfaction does not seem to matter. The smiling faces and so many other trivial scenes that train AI make it possible to maintain the violence described by Schiavinatto. Putting these images in a historical perspective and creating experimentations to rethink the practices behind them is a way of reacting to the problems that computer vision maintains and updates.

The five experimentations discussed here also indicate that computer vision is so intertwined with colonial practices of Western scientific 'culture' that the inclusion of more diversified data and the improvement of the working conditions of those who train AI do not seem enough for a less problematic computer vision. The outputs obtained through the experimentations discussed show the urgency of a refoundation of this AI subfield. Creative and artistic experiences can help in the construction of more radical possibilities for the way computers could 'see' our world. In this way, the images that train machines and the processes made possible by them can finally be part of a public sphere of discussion – and not just compacted and hidden in folders and subfolders of datasets.

Acknowledgements. Richard Staley, Sarah Dillon, Jonnie Penn, participants of Histories of AI: A Genealogy of Power (University of Cambridge), Fabio Gagliardi Cozman, C4AI Inova USP, Group on Art and Artificial Intelligence (GAIA), Lucas Nunes, Bernardo Fontes, Gabriel Pereira, Mariana Mendes, Giselle Beiguelman, Faculty of Architecture and Urbanism (University of São Paulo) (particularly students and professors of the *demonumenta* project), Anand, Pedro and Sonia.

61 Personal conversation, 2 November 2020.

Cite this article: Moreschi B (2023). Five experimentations in computer vision: seeing (through) images from Large Scale Vision Datasets. *BJHS Themes* **8**, 171–187. https://doi.org/10.1017/bjt.2023.6