AI & SOCIETY
https://doi.org/10.1007/s00146-020-01093-w

ORIGINAL ARTICLE

On the data set's ruins

Nicolas Malevé[1]

Received: 5 August 2020 / Accepted: 14 October 2020

Abstract
Computer vision aims to produce an understanding of digital image's content and the generation or transformation of images through software. Today, a significant amount of computer vision algorithms rely on techniques of machine learning which require large amounts of data assembled in collections, or named data sets. To build these data sets a large population of precarious workers label and classify photographs around the clock at high speed. For computers to learn how to see, a scale articulates macro and micro dimensions: the millions of images culled from the internet with the few milliseconds given to the workers to perform a task for which they are paid a few cents. This paper engages in details with the production of this scale and the labour it relies on: its elaboration. This elaboration does not only require hands and retinas, it also crucially zes mobilises the photographic apparatus. To understand the specific character of the scale created by computer vision scientists, the paper compares it with a previous enterprise of scaling, Malraux's *Le Musée Imaginaire*, where photography was used as a device to undo the boundaries of the museum's collection and open it to an unlimited access to the world's visual production. Drawing on Douglas Crimp's argument that the "musée imaginaire", a hyperbole of the museum, relied simultaneously on the active role of the photographic apparatus for its existence and on its negation, the paper identifies a similar problem in computer vision's understanding of photography. The double dismissal of the role played by the workers and the agency of the photographic apparatus in the elaboration of computer vision foreground the inherent fragility of the edifice of machine vision and a necessary rethinking of its scale.

Keywords Computer vision · Data set · Photography · Imageability · Scale · Micro-labour · Le musée imaginaire · ImageNet

1 Introduction

In *On the Museum's Ruins*, Douglas Crimp (1980), reflects on the proximity between the museum and the mausoleum. The museum is the place artworks, disconnected from the practices and lived conditions that breathe life into them, have come to die. Crimp's text dramatises a moment of change in the museum's history. The museum, once considered the shelter of a selection of significant artworks, evolves into a site, where genres that were deemed secondary are excavated from the storerooms and given exposure. The inclusion of salon painting in the installation of nineteenth century painting at the Metropolitan Museum signals, for Crimp, the demise of modernism. And it triggers, among his contemporary critics, a wave of criticism against the lack of rigour, the laissez-faire attitude of a postmodern era. Refusing to embrace such criticism and its nostalgia, Crimp moves on to articulate a distinct critique of the new museal institution. The pretension to knowledge from the part of the museum is to obtain a coherence from objects and works extracted from increasingly heterogeneous sources with art history being assigned the role of homogenising the disparate museum's contents. To extract knowledge through a dialogue between regularised fragments is the condition at which the museum resists its assimilation to the mausoleum. Crimp sees this tendency to a larger inclusion expressed by André Malraux's *Le Musée Imaginaire* in which oeuvres from all civilisations can be confronted and compared. As Crimp remarks, Malraux's project emphasises the fact that art history cannot produce a universal plane of comparison without another organising device: photography. Photography, here, acts as a leveller. Every art object that can be photographed can enter Malraux's museum. Through the

✉ Nicolas Malevé
maleven@lsbu.ac.uk

[1] Centre for the Study of the Networked Image, London South Bank University, 103 Borough Road, London SE1 0AA, UK

Published online: 11 November 2020

photographic medium, a part of a plate can be compared to a detail of a sculpture, an enlarged brush stroke to a mural painting. Photography is a condition for the inclusion of objects and for the museum's knowledge production. But to function this way, photography must be made a transparent vehicle. As soon as photography itself is considered as an autonomous agent, heterogeneity comes back and the museum's pretension to knowledge is in jeopardy. The ruins, for Crimp, are not only those of the museum's contents which roots are severed from the soil that gave them birth. The ruins are those of the museum itself, an institution of knowledge production based on the active yet transparent role of a levelling device: the photographic apparatus and the scale it introduces. Repressing the apparatus through which it obtains its coherence makes it blind to the artificiality and the limits of the knowledge it produces.

The current text is not an essay on art history. In the following pages, I will address the problems of a discipline of computer science, computer vision (CV), in which engineers train machines to make sense of images and, for that purpose, assemble huge collections of photographs. In this endeavour, I won't use Crimp's critical dialogue with Malraux as a strict analytical framework, but rather as an insistence and call to increase sensitivity to the relation between the photographic device, scale and a pretension to knowledge in the vast sets of images assembled in computer vision. It is with the notion of selective use of photography, and its repression, related to the homogenisation of a potentially limitless source of visual imagery, that I interrogate the formation of computer vision. To understand the relevance of such a conceptual device for CV, it will be necessary to take a step by step approach to photographic practices at the core of the discipline.

2 Computer vision, learning to see by example

Nowadays, contemporary computer vision is increasingly understood as a sub-field of artificial intelligence and machine learning. This is the disciplinary affiliation given in introductory courses and manuals (Brownlee 2019; Fullscale 2019; Wikipedia Computer vision no date). To construe CV first and foremost as a field of artificial intelligence (AI) means several hierarchies are taken for granted. For instance, it means that a discipline, AI, applies the same techniques, with a few variations, to different kinds of data. In computer vision, visual input is often unproblematically conflated with photographs or media abiding to the representational codes of the photographic image.[1] Photographs understood as data are presented as passive samples awaiting the mining of algorithms to be made meaningfully part of computational systems. In such a view, agency is located on the side of the algorithm, and data, as the name suggests, is simply given.

The work of Fei-Fei Li, leading computer vision scientist, director of Stanford AI, and initiator of the ImageNet data set, shows how a different valence can be given to photographs. As she (GoogleTechTalks 2011) states, "another way to understand CV is through the evolution of its data sets". Taken seriously, data sets challenge an algorithm-centric view of CV. The algorithm becomes relative. For Li, data sets are no longer suppliers of data, they are tools to formulate CV's problems: they engage actively in the modelling process. To understand such a claim, we need to look at how modelling functions in contemporary machine learning systems.

Traditional AI and early computer vision relied on explicit modelling. In a framework, where modelling is explicit, developers design themselves a model that matches the complexity of the problem domain. For instance, a cat's face is decomposed into simple shapes like a circle for the face, two triangles for the ears and two circles for the eyes. This approach is only efficient for a limited amount of cases. Its advantage is to produce a legible model: circle + 2 triangles + 2 circles = a cat face. Such a model may function in a strictly controlled environment, but leads to overwhelming problems in real-world scenarios, where the cat may appear from profile, with eyes closed, at rest or jumping. Additionally, one cannot expect the animal to be perfectly centred and illuminated and partial occlusions or unexpected perspectives often change the organisation of the patterns. As neuroscientist David Marr remarked, photographs rarely follow a predictable order. Noticing the almost despairing feeling of early computer vision researchers, he concluded that "practically anything could happen in an image and furthermore that practically everything did" (Marr 1982, p. 16). Every time a photograph deviates from an expected pattern, the algorithm needs to be updated and optimised. Such an approach lacks generalising power as the algorithm must not merely discriminate a pattern adequately, it also needs to make sense of the other patterns interfering with what it attempts to detect (Deng et al. 2010). Furthermore, the developer needs to learn in detail about the object to detect and to analytically decompose it before writing the code. Under this paradigm, to write a cat classifier, a programmer needs to become a feline expert.

[1] See, for instance, one of the most cited papers in the discipline, "A large-scale hierarchical image database – ImageNet" (2009), in which image and photograph are used interchangeably.

In contrast, current techniques of machine learning based on neural networks do not rely on a previous analytical decomposition. The developer assembles a data set reflecting the variations of the domain under study and utilises automated means to calculate an optimal function that treats the features of the data as parameters. In computer vision, this technique, at its most simple level, uses large visual databases in which discrete units such as pixels can be considered as data points. Common techniques of machine learning in computer vision are said to be "supervised", which means that the data is curated (Beheshti et al. 2016) to provide examples from which the machine learning algorithm extracts regularities: the software "learns by example". To come back to the case of the cat, in the data-oriented paradigm, the developer does not try to decompose the animal in distinct shapes and explicitly summarise their relations. Instead, she curates a large series of photographs, where the cat is displayed in various positions, and lets the algorithm detect the regularities traversing the various samples. Through this phase of "learning", the algorithm produces the model of the cat.

The change in algorithmic design does not merely correspond to an evolution of the technical process. It provokes more largely a reconfiguration of the positions of the actors in the field. The analytical decomposition of the problem is now replaced by the production of a data set exhibiting the regularities that define the problem domain. Engineers are said to write programs that "discover" the rule inherent to the data (Simpson Center for the Humanities UW 2017). However, these data sets do not appear by magic, they need to be curated, assembled, maintained and annotated. Concretely, the modelling is outsourced to those who curate and annotate the data set. The annotators and curators are in fact implicitly coding the model that will be discovered algorithmically. The engineers say the model is learned end to end. This means in fact that it doesn't learn from them anymore. In the current machine learning paradigm, the engineer doesn't need to be a feline expert to produce a cat detector, but the engineer relies on a population of curators and annotators actively engaged in defining what counts as photographs of cats. The paradigm change in machine learning has externalised the modelling process and, by doing so, has produced a new division of labour.

If we take this seriously, then the data set becomes a site, where computer vision is made, rather than a component of its supply chain. The techniques to compose a data set are not subsidiary to the problem of computational vision, they are integral to it. But as important as data sets are for machine vision, two intertwined aspects of their production are generally overlooked: the micro-labour of annotation and the role given to photography in the process. To examine the labour that goes into CV and its photographic dimension, I will analyze how ImageNet, one of the largest database of human annotated visual content to date (Deng et al. 2009) has been assembled. The ImageNet project is a collection of images for visual research that offers tens of millions of images manually annotated, sorted and organised according to a taxonomy. It aims to serve the needs of computer vision researchers and developers for training data. Due to its size[2] and its extensive annotations, ImageNet has become de facto the most used knowledge base in the world of computer vision (Fei-Fei 2010). Following the work of the annotators who selected the photographs and reflecting on the role given to photography, I will try to understand their roles in the process and the reasons why they remain treated as contingent rather than crucial to the definition of what it means for machines to see.

3 The elaboration of CV

Computer vision data sets depend on a standing reserve of large volumes of manually annotated photographs. Advertising a new update of the Google Photo service, Chuck Rosenberg (2013) wrote that, thanks to the advances of computer vision, users could search their own images *without having to manually label each and every one of them*. This tedious work can be executed automatically from now on. Yet, to automate this task, large volumes of annotations are themselves necessary. While the users of the new Google software may not be obliged to manually tag their photos anymore, thousands of hands on keyboards, retinas glued on screens are producing an unprecedented amount of labelling and descriptions to train the algorithms that automate the user's tagging. Perhaps, the amount of annotation work involved in the production of data sets is more impressive than the amount of photographs included in ImageNet. After all, 14 million images are only a fraction of the monthly 575 million public uploads of photos on a platform like Flickr. (Smith 2019) The work of manually cross-referencing and labelling the photos is what makes data sets like ImageNet so unique. In fact, there has been rarely in history so many people paid to look at images and report what they see in them (Vijayanarasimhan and Grauman 2009). The automation of vision has not reduced the amount of eyeballs looking at images, of hands typing descriptions, of taggers and annotators. On the contrary, it has increased their number. Yet what has changed is the context in which the activity of seeing is taking place, how retinas are entangled in heavily technical environments and how the speed at which candidate images must be evaluated. Due to the pressure of generating annotations at the scale of the web, the process

[2] The average number per categories is 10.5k, see ImageNet_2010.pdf.

needs to be massively "parallelized".[3] A parallel architecture optimises the calculation and makes maximal use of all the computational resources of a machine. On the Amazon Mechanical Turk (AMT) platform, where a large part of the annotation effort is produced, the workers are treated as computational processes. The workers, the "Turkers", are abstracted away. The advantage of having parallel processes is that the execution time of a given task can be divided among the available workers/processes. Li calculated the amount of human labour required for ImageNet this way: estimating that a person can annotate two images per second, such that 19 human years are necessary for an individual working 24 h a day to verify the tens of millions of images in her data set. AMT allows her to compress the 19 years into 2 years by providing the necessary workforce to handle the workload in parallel (GoogleTechTalks 2011).

Apart from speed, there is another less "technical" reason to adopt such architecture. The parallel architecture isolates the workers and makes it difficult for them to create bounds, to build a collective identity and to unionise. The Turkers have no name, only an anonymous identifier and the platform doesn't provide any mean of communication between them (Irani and Silberman, 2013). There are political reasons for crowdsourcing platforms to fear workers' unions: their working conditions are exploitative. AMT offers people with large data sets the possibility to outsource the annotation work using their massive workforce. The tasks on AMT are rewarded with micro-payments. An annotation is estimated between 1 and 4 cents and the estimated hourly revenue for a Turker is approximately two dollars (Hara et al. 2018). The annotator must find an optimal trade-off between the precision and attention required to make a good enough annotation and a speed that allows her to "maximise" her financial gain. The hourly rate depends on the task and access to the most lucrative tasks depends on the age, the origin and the class of the workers as one needs to be "culturally compatible" (Irani 2015, p. 726) with the request. Annotation tasks in particular require to quickly learn the knowledge relevant to very different subjects including religious differences (i.e., be able to discriminate between photographs of Catholics and Old Catholics, archbishops and archpriests, Buddhist and Zen Buddhists from sources as various as the Vatican website, English tabloids, and family albums), military ranks and uniforms (i.e., differentiate between Redcaps and Green Berets, adjutant generals and generals, sergeants and first sergeants, recruiting sergeants and gunnery sergeants, from sources ranging from military websites, news outlets, TV series or wedding pictures), or bacterial species (i.e., identify the *Bacillus anthracis*, the spirillum, the clostridium perfringens, or the gonococcus from microscopic imagery).

The workers, if they want to make a living, need to work at a pace that barely allows them to see the images. For the annotators, structurally, the "glance" is the norm. Speed is built in the platform economically. Training sets must be produced fast. Lots of workers are mobilised intensely for a short period of time. Through the interface of the AMT, the requesters are managing the cadence of the annotation work. They want to ensure the workers go fast enough to match production deadlines. And, at the same time, they attempt to preclude them from overlooking their task to avoid drops in quality. The interfaces of annotation are designed to control workers' productivity, to find the optimal trade-off between speed and precision. The time estimation for labelling tasks is especially difficult as it varies according to the annotator's experience and the nature of the photographs. Andrej Karpathy, now Tesla's director of AI, reported that when he annotated an ImageNet's recognition challenge's data set, the labelling started at a rate of one image per minute and decreased with the accumulated experience (Karpathy 2014). But, even if progress was noticeable, the rate was not constant as some images like those depicting some particular breeds of animals required a longer time and additional research. To cope with the variability of the annotation process, numerous techniques are currently tested to streamline it. To begin, as manual labelling is costly, the priority must be given to the annotations producing the highest information gain. As Vijayanarasimhan and Grauman put it, unlabelled images must be ranked "according to their expected 'net worth' to an object recognition system" (Vijayanarasimhan and Grauman 2009). The cost of the labelling effort leads the computer vision researchers to approach visual content in the form of informational currency and attention scarcity. This approach informs the architecture of annotation workflows, wherein a pre-labelling attempt is made by an algorithmic detector and corrected by human annotators (Papadopoulos et al. 2016). Pattern recognition techniques are used to absorb the bulk of the "easy" detection work, spot the potential targets and redirect the "ambiguous" cases to the human annotators (Vijayanarasimhan and Grauman 2009). While these techniques of semi-automation are in their early stages, their existence indicates the anxiety of the data set makers to control the annotator's attention and prevent her distraction. The requesters invest in the annotator's attention and treat attention as an asset that needs to be protected. Besides guiding the annotator's eyes to specific regions of interest and regulating their attention, researchers are exploring how to augment the annotators' ability to absorb visual content. For instance, Krishna et al. (2016, p. 2) claim to have devised a technique of rapid serial visual

[3] In Informatics jargon, to "parallelize" a process means to design a workflow where two processes can take place without interference.

presentation that allows workers to produce one "hit"[4] in 100 ms by immersing them in an uninterrupted visual flow. As the volume of requests augments, such experiments indicate that the unit of measurement for hits is moving towards the millisecond. Finally, others are concentrating on evaluating workers' performances over large periods to identify the annotators able to sustain the rhythm over time without decline in submission quality and sifting out the "satisficers" (Hata et al. 2017) who strive to do the minimal amount of work to meet the acceptance threshold. In response, the workers themselves try to figure out what is a good ratio between speed and accuracy. On social media platforms like Reddit, AMT workers exchange tips about "good paying tasks" and compare their performances. To be able to estimate the amount of work required for a task is a matter of survival in the crowdsourcing environment: both for the data set maker who is under pressure to deliver in a competitive environment and the annotators who internalise the system's speed.

Where does this analysis of the annotation architecture leave us? First, it emphasises that data is not given. Data needs to be produced and engineers are deeply involved in setting up the material conditions of production of said data. If the work of curating data sets and their annotation is central, so is the management of the populations of workers involved. More importantly perhaps, we see how the notion of scale transpires into the whole process. With this notion of scale, we find a point in the midst of the technicalities of CV, where the discussion of *Le Musée Imaginaire* starts to resonate. A scale here is not a simple measurement of the accumulation of more material which has for consequence an accumulation of knowledge. Scaling corresponds to a production of difference not just to an increase of the same. If Malraux emphasises that his contemporaries have access to more artworks than their predecessors, he also insists that their relation to art differs in kind as well as in breadth. They have access to more categories of works and to a larger selection of works inside these categories. The expansion of categories and expansion of intra-categorial comparisons are the engines of a different relation to artworks and their representations. In this perspective, *Le Musée Imaginaire* is more than a printed book, it is an operation that brings the museum to the scale of the printed press.

Even if Li's concerns feel remote from Malraux's concerns, to read *Le Musée Imaginaire* from this perspective, invites to pay closer attention to the generative character of scale. It shifts our attention from quantity increase (the addition of discrete content) to scale as a qualitative architecture of relations. And it helps take the measure of what is at stake when Li declares ImageNet is meant to bring computer vision to the scale of the web. Between the French Minister of Culture and the director of the Stanford AI lab, there is a resonance in the investment in scale as vector of knowledge transformation. There are also striking differences. For instance, in Malraux's essay, one could hardly find any reference to the labour involved in the production of the photographs or any economical consideration. Malraux's museum is abstracted away from these concerns, as the art historian benefited from the budgets and apparatus of the Ministry of Culture. Li, on the other hand, brings the logistics and the economics of scale to the front. When she states that a crucial step to advance her discipline is to resolve the scale of the problem, she makes clear that this involves getting one's hands dirty with issues of management and control.

This dissonance calls for further elaboration on the nature and dynamics of the knowledge that is produced. A scale, manifested in logistics and industrial infrastructure, is deeply generative. As media scholar Matthew Fuller writes, a scale provides "a certain perspectival optics by which dimensions of relationality and other scales may be 'read' […]" (Fuller 2005). From a newly sensible scale, it becomes possible to "read new dimensions of potentiality" (Fuller 2005). Li's work is an articulation of scales. It is the millions of images with their labels, the articulation of vision with the decomposition of work in micro-tasks and micro-payments. The problem of scale is an articulation of both the macro and the micro. 14 million images need a model of vision, where the milliseconds are the unit of measure for perception. When thinking about the problem of scale of computer vision, we need to keep in mind the opposite poles of the problem's dimension: on one hand, the infinitesimally small units of perception (the eye saccades), the miniaturisation of the work process in discrete hits that can be performed at full speed; and on the other hand, the massive amounts of photographs available on digital platforms and the vast population of precarious workers ready to annotate on demand.

This articulation of scale brings about a new distribution of labour crucially affecting the task of modelling. A task that was previously in the hands of the engineers is distributed among the annotators. The engineers delegate a series of crucial decisions, while keeping control over the way the problem is formulated and its interpretation. Through the interfaces and contracts of AMT, the engineers design the frame of "transepistemic relations" (Knorr-Cetina and Malkay 1983).[5] For Karin Knorr-Cetina, an example of a transepistemic relation can be found in the interaction between a funding agency and the researchers it supports. Often the

[4] In Amazon Mechanical Turk's jargon, a "hit" stands for a "human intelligence task".

[5] "Transepistemic", an adjective coined by sociologist of science Karin Knorr-Cetina refers to relations crucial to the inquiry that exceed the boundaries of the scientific community.

funding agency participates to the framing of the research problem by negotiating the methods to be used or the interpretation of measurements (Knorr-Cetina and Malkay 1983, p. 132). Research happens in and ex situ, yet the contribution of the agency remains unacknowledged in the reports and research outputs. To produce "pure" knowledge, the scientists do not refer to the funding agency as a collaborator as it would open the door to a criticism of external influence, and a suspicion of introducing non-scientific criteria. With this concept, Knorr-Cetina emphasises the active role of external agents in the production of science and their relative invisibility. The concept stresses the importance of the circulation of knowledge and its distributedness. Although in an inverse balance of power, the engineers and the annotators together contribute to the production of the knowledge relevant to computer vision. AMT can be characterised as a transepistemic device that enables the collaboration while masking one party's contribution. It elaborates computer vision as it provides the modelling for the training process at the appropriate scale, and it elaborates it (this time with the emphasis on *laborare*, the latin root of the word) as it articulates computer vision's division of labour.

4 The data set as photographic alignment

What is the nature of the workers contribution in the transepistemic relation? Workers are said to be *cleaning* the data set, a term that invokes hygiene and manual rather than intellectual labour. To clean is to scrub away the germs and parasites intoxicating a set of samples. It also indicates clearly, where is the annotator's side in the division of labour. By contrast, the role of the worker would be framed much differently if she was considered as contributing to the curation rather than the cleaning of the data set. The nature of the decisions the annotator has to take is of central importance for what is included in, or excluded from, the data set. Yet, as we begin to understand better, the labour so crucial to the process remains undervalued financially and unacknowledged (the workers are anonymised and interchangeable). Furthermore, the decision-making power of the workers, as we will see, must be simultaneously enabled and delimited.

With the annotator's work, we are at the core of the process of regularisation of computer vision's objects. Regularisation means to process the objects in a manner that makes them suitable for use in scientific work. In this case, it means to fashion the objects in a manner that they exhibit the regularities that will be picked up by the algorithms at a later stage. As the historians of science Daston and Gallison (2017, pp. 19–22) remark, "No science can do without such standardized working objects, for unrefined natural objects are too quirkily particular to cooperate in generalizations and comparisons." Regularities are never simply given in the data, they always need a helping hand. The workers are given a set of photographs pulled from the Internet through automatic queries of visual search engines. The results from these queries are noisy, ambiguous, and at times obviously unrelated.[6] A considerable amount of work of disambiguation has to be performed. The workers are tasked to filter the search results and extract the relevant photographs. They are expected to find by consensus which photographs are "imaging" a given concept. Imageability, a concept imported from psycholinguistics, describes the ease or difficulty with which a concept can be characterised visually (Paivio 1986; Yang et al. 2019). ImageNet is composed of a wide array of heterogeneous entities including stars, cities, animals, vehicles, micro-particles and diseases. But as different as they are, computer scientists think they may all be *imaged* in a way that make them available for comparison and differentiation. Imprints of light captured by a digital camera, computer simulations, screenshots or photoshoped images are all reduced to pixels. Therefore, molecules, pencils, acrobats, coliphages, olives and tetraspores belong to a same *imaged* register, where regularities and differences can be observed. What do they have in common? They can be photographed, or more precisely represented according to the codes of photographic realism. Significantly, the interface warns the ImageNet's annotators: "PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc." A pivotal role is given to photography conceived as a leveller, an instrument that automatically converts light into pixels according to predictable rules and at the same time that *images* a concept. Photography is mobilised as an instrument to homogenise the visual world, to transform the visual into data, where data of different origins can be compared and classified.

In the first part of this text I have discussed how the notion of scale was central to Li's project. I have differentiated scale from a simple quantitative indicator. Resolving the scale mobilises micro-labour, eye saccades, and billions of data items. Now we see that the billions of elements selected to enter the data set are not random visual files, they are photographs, and as photographs, they are treated as suppliers of representations, dedicated to the task of "imaging" concepts. Furthermore, photography is given a role of leveller which offers a plane over which individual representations can be compared. In *Le Musée Imaginaire*, Malraux forges a word that encapsulates photography's role in regulating the representations entering the collection. Everything in the museum has a place if it is "photographable" (Malraux 1965, p. 123). By photographable, Malraux doesn't mean only something that

[6] Referring to the Torralba et al. (2008), Yang et al. evaluate to 10% the number of photos matching a query, https://people.csail.mit.edu/torralba/publications/80millionImages.pdf.

can be reproduced as a photograph. In Malraux's account, photography is active and is perfectly suited to address the question of style. For him, a photograph doesn't transparently describe an artwork. Light, frame, composition, angle, all contribute to reveal style within the visual object. Yet even if these correspond to carefully selected choices, they are revealing something that exists within the object. The photograph makes the objects speak. This revealing, even if it is obtained through a technical procedure, is by no means purely mechanical. In the book, Malraux compares different photographic reproductions of a same artwork and insists some capture the artwork better than others. Therefore, Malraux's photograph is not just a passive reproduction. However, as a revealing, its active contribution is intimately bound to its effacement, its disappearance. Any surplus to what it reveals can be dismissed as irrelevant, and when two photographs are juxtaposed (or "confronted" in Malraux's terms), the dialogue we are invited to witness happens between two objects, not between two photos. To be photographable means to be opened up to comparison by the means of photography.

To bring the discussion back to computer vision, the problem faced by the data set creators can be formulated this way: how to make 20,000 categories imageable and photographable? Diversity is an important criteria to constitute a data set. The diversity in question concerns the objects represented in the photographs. The objects must exhibit variance in positions, viewpoints, appearance. They must be placed in front of different backgrounds and must appear with various degrees of occlusion (Deng et al. 2009, pp. 4–5). The search engine is the provider of such diversity as it concentrates in one page, photographs of distinct sources. Concretely, it means the engineers delegate to the search engine the task to extract the photos from the context in which they operate online. To perform this extraction, the search engine needs to undo a large series of relations that held these photos in place. For example, as most ImageNet's photographs are originally posted on Flickr, it requires the undoing of the relations the platform has established among the many entities that partake in the photograph. On the platform, the photograph is a composite. It is made of a series of JPEG files of different formats. It includes tags and comments. It circulates in communities and albums. Metadata information are attached to it, and so on. All these relations form an "alignment" that makes the Flickr photograph an entity that can be shared, liked, viewed rated. In the Flickr ecosystem, the photograph has currency. When the photograph gains popularity, it opens the door to new groups, the photographer receives a badge, and his visibility increases. The term alignment stresses the fact that a photograph doesn't exist alone. It is enacted through a series of relations taking place in an apparatus (in this case, the Flickr platform) through which it gains properties, affordances, and currency.

To be included in the data set, a digital file is excised from its Flickr alignment. The comments do not travel with the file, and neither do the albums, metadata information, or the author's name. The Flickr photograph, once tagged by the author or the community, is now categorised according to the WordNet thesaurus. This operation is far from innocuous. The photograph is enacted differently. It is re-aligned, and this re-alignment requires a considerable amount of work. Once selected by the search engine, the files are acquired by the data set makers and they are shown in grids of hundreds of thumbnails to the Turkers. Moving from the original context, where they were published, they enter into another framework of attention. With the AMT interface, a photograph once seen in an individual page or inside a blog article is displayed among many other candidates. The screen is filled with thumbnails. Proximities change radically, the photograph sits next to new neighbours. As already mentioned, AMT privileges the glance rather than a sustained observation as the exploitative labour conditions imply a rapid pace. Decisions must be taken fast. Moving from Flickr to AMT is not just a change in location and relations, it is a change of rhythm, speed, a change of metrics. It is also the conversion of an economy of sight. The Flickr photograph, a product of free labour, becomes the AMT thumbnail, an object of minimal attention for the Turker struggling to make ends meet. The translation from free labour to micro-payment goes along with the translation from an environment that celebrates the full screen view to a device that optimises its workflow with the thumbnail. In AMT, considerations of aesthetics do not apply, legibility becomes more important. Imageability is the dominant criteria of selection and the worker's competence is her ability to capture the photograph's content fast.

5 Decision-making, micro-labour in the photographic alignment

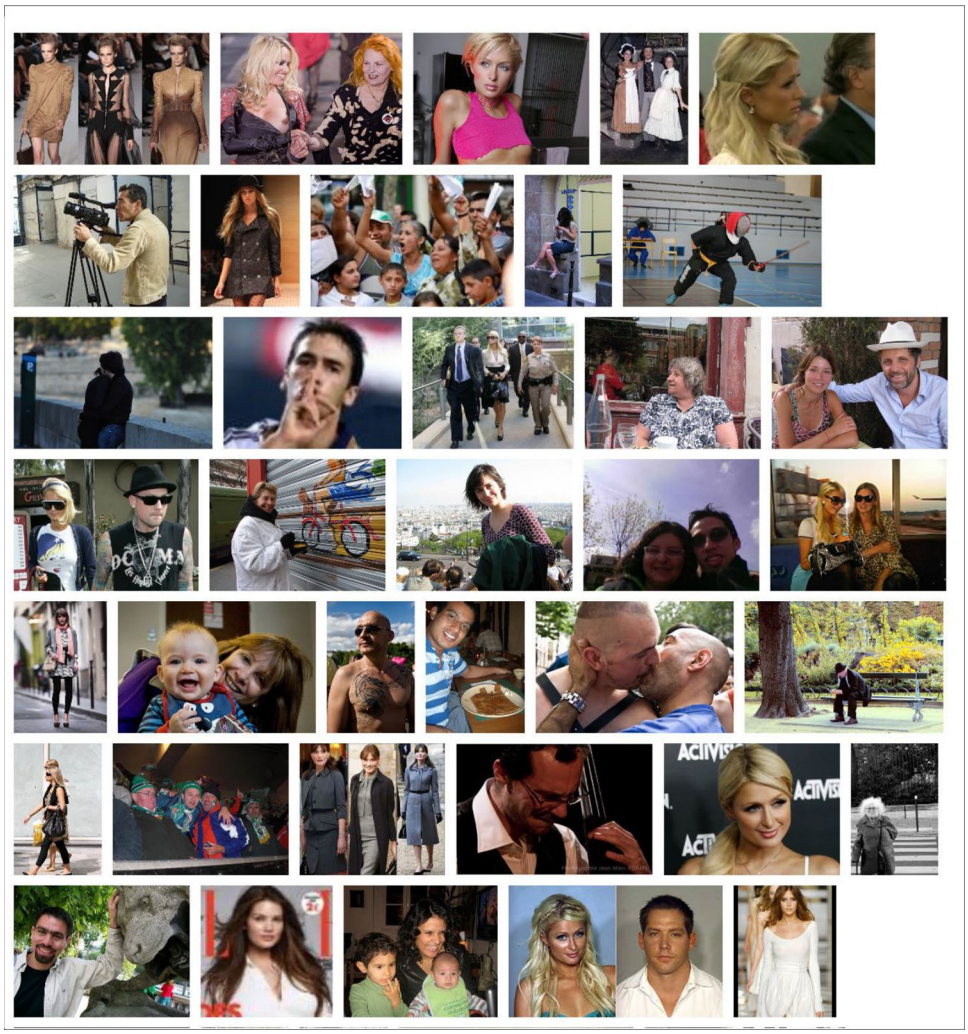
A series of examples will help understand the imaging process happening through the re-alignment of the search results. I will begin with the synset "Ratatouille". The description given to the workers for the ratatouille concept is "a vegetable stew; usually made with tomatoes, eggplant, zucchini, peppers, onion, and seasonings". It is filed under Misc → food, nutrient → nutriment, nourishment → dish → stew. The photos selected in the synset are for one half extracted from Flickr accounts and the other half from culinary websites, foodie blogs, cooking tutorials, or restaurant pages. The distribution of photographs in the synset seems to correspond to the stated goals of the

Illustration 1 synset n07592768, Ratatouille

data set makers: variations in shapes, lighting, and points of view. The selection is testament to the variety of forms the dish can take. Layered slices of vegetables in a baking tray, cubes of tomatoes and eggplants in a pot, the meal is displayed alone or in combination with other dishes, on a stove, next to a sink, on a garden table at a barbecue party, in a living room on an immaculate white sofa, in a tupperware container on a flower tablecloth. The technical competences of the photographers vary. The framing and cropping of the photographs, the sharpness, the blurs, the light balance relate to different intentions and belong to contrasting aesthetics. A photograph capturing the delicate reflections of light in the marinade sits next to a quick snapshot in which the details of the meal vanish under reflection of the flashlight. In many ways, the variety of the selections meet the engineering requirements. Yet a closer look at the selected items reveals the presence of entities that do not match the synset's definition. The dish has been confused with various forms of stew (e.g., meat stew) or tomato and mozzarella salads. More surprisingly, pasta meals and grilled shrimps on skewers have also been confused with ratatouille. The Pixar character of the eponymous movie also appears to be a candidate for the combination of zucchini, tomatoes and eggplants. These items can be seen as the results of mistakes by individual workers who didn't pay enough attention, and that can be partially explained by the nature of the dish inasmuch as it takes multiple forms (e.g., a "rata" in French military slang originally being a rough mix of cheap vegetables) (Illustration 1).

The data set makers anticipated human errors of judgement and the difficulty to decide whether the candidate images contain a particular object. They designed a solution to evaluate inter-annotator agreement. Essentially, a candidate image is included in the data set only if a consensus is reached among annotators. As I have explained, the platform relies on parallelism for technical and organisational reasons. The workers are never aware of the decisions taken by others and cannot interact with them. Consensus on the platform means that several workers are presented the same candidate images and the platform compares their selection.

Illustration 2 synset n09708750, Parisian



Where the selections overlap a consensus is reached. Furthermore, this consensus is modelled according to a scale that varies according to the "semantic difficulty" (Deng et al. 2009) of the data set (its grade of imageability). The ImageNet authors give the example of the difficulty to reach a consensus for a synset like "Burmese cat" in comparison with "cat". As cat is deemed more imageable than Burmese cat, the number of annotators who need to agree on the label "Burmese cat" for the same photograph needs to be higher than the number of agreeing annotators for "cat". "Ratatouille" at the bottom of the branch Misc → Food Nutrient, is one of those terms deemed more semantically difficult and requiring, therefore, more eyeballs and a higher level of consensus.

Yet a higher level of consensus doesn't guarantee a greater conformity to the definitions given to the workers. In the synset n09708750, the concept "Parisian", is described as "a native or resident of Paris" (Illustration 2). A remarkably high portion of the selected images are photos of Paris Hilton in different forms including candid and tabloid photos, 3D models, or selfies. Representations of the socialite in bathrobe, bikini, gown or casual wear dominate by far other familiar cultural tropes of Parisians like tourists at the Louvre Pyramid, passers-by wearing berets and striped sailor shirts, men kissing at Gay Pride, people enjoying a coffee at a terrace, or visiting the Eiffel tower. Unlike the previous example, the imbalanced proportion of photographs of Hilton cannot be explained by the errors of a minority of annotators that have introduced a negligible percentage of erroneous candidates.

The cause is often attributed to the workers. Engineers develop techniques to track the workers that are considered as "satisficers" (Hata et al. 2017) or "spammers" (Quach 2019) who accomplish their tasks with too little care. Other causes are evoked, such as the annotators are not culturally compatible (Hata et al. 2017) with the request or they don't read the definition. For these reasons, many precautions are in place to ensure the definition is read by the annotators. Before seeing the candidate images,

the annotator is presented the definition of the synset. Subsequently she is presented another screen, where she has to choose the right description among several others. For instance for the synset N04070003, "reformer", the annotators are given the following "gloss"[7]: "An apparatus that reforms the molecular structure of hydrocarbons to produce richer fuel; a catalytic reformer" (imagenet.org no date). However, none of the photos selected in the synset depicts such an apparatus. Instead they depict Pilates reformers and people doing exercises, even if to access the hit, the annotator had to confirm explicitly that reformer was a chemistry-related apparatus and not a gymclass prop. For the annotator, glancing is not only a mode of perception related to the rapid scanning of thumbnails, text too is read in a glimpse (Illustration 3).

This brings us back to the problem of decision-making in the annotation environment. The decision is delegated and regulated through consensus. Supervision happens *post-hoc* when consensus cannot be achieved or through punctual quality checks. Supervision here differs starkly from an idea of a subject dominating a scene from a bird's eye perspective. To contrast, it may be useful to remind ourselves of a famous photograph representing Malraux in his spacious office at the French Ministry of Culture. The floor is covered with photographs. Malraux standing in front of his desk holds one of the photographs for further consideration. Malraux literally supervises, visually dominates the pictures distributed at his feet. This mode of supervision does not correspond to any of the actors involved in ImageNet. Li and her colleagues control the process through the reports given by the AMT interface. They receive numerical indicators and they can insert them in spreadsheets. But they never supervise ImageNet's visual content from an overhanging position. There is no floor large enough to contain 14 million photographs, and even if such floor could exist, there would be no position from which an observer could embrace its totality. If such a position is not available to the researchers, the situation is even less comparable for the annotators. If the annotators are given the role to look at the photos on their screens, they are immersed in a flow and unable to negligently pick one photograph and consider it with an air of detached interest. This doesn't mean, however, that no decisions are taken and that no supervision is in place. What this comparison suggests is that we need to further enquire into the mode of decision-making and that we have to renew our notion of supervision to adapt to what is happening in the annotation environment.

To annotate at speed does not consist of a mechanical response issued from a passive subject. To understand the annotator's contribution, it is fundamental to understand the process as one of elaboration which goes beyond rational choice and explicit judgement. The epistemic contribution consists in embodying a scale, figuring out rhythms and levels, understanding and refraining involvement. To attend to the process of elaboration means to avoid concentrating exclusively on the semantic decision. The elaboration is not limited to a pivotal moment, where the annotators assert the meaning of a photograph. It includes the complex methods through which they synchronise within an alignment and embody a scale. Synchronisation, scale embodiment, attunement, and seeing at speed are at the core of the photographic mediation of computer vision. They intervene crucially in the resolution of the photograph in a given alignment. To understand the annotator's contribution is, therefore, to be attentive to how she creatively relates with the apparatus, how she probes, where the apparatus begins and ends. To attend to the resolution of a candidate image into a data set item requires a rethinking of the nature of the decisions that are taken. It requires to shift from an understanding of a decision mechanism based on a pivotal moment that involves an "isolated who" (Mol 2002) to a more distributed consensus of actors and devices. The question becomes: how do the Turkers achieve consensus without explicit coordination? A crucial part of the answer lies in the methods the annotators mobilise to figure out the rhythm they have to follow.

The cadence of the platform comes from the remuneration. To secure a minimal income, the Turkers have to perform at high speed. They have to calibrate their involvement. They need to figure out how precise they must be to make enough annotations and have their hits accepted. This balancing act requires finding one's place in the alignment. The Turkers are contributing to the resolution of a scale, to the correlation between semantic hierarchies and speed. They do not debate together to decide whether a candidate image should be considered "tomato mozzarella" and consequently be excluded from the "ratatouille" set. They have to judge whether it is worth slowing down to give attention to this candidate or if, at first glance, it can be assimilated to the concept ratatouille without threatening their remuneration. They have to intuit if the difference they notice is worth changing pace or if they can remain indifferent to this difference. The Turkers develop a common sense of the apparatus' resolution without having to explicitly negotiate. The architecture and the pace of the AMT define its resolution, its grade of precision in the descriptions, and concomitantly, it creates a zone of indiscernibility for the apparatus.

A consensus is not always a matter of explicit argumentative deliberation. It can be a matter of levelling, of finding a common grade of involvement. To ask which arguments led to the selection of some items rather than others is less important than asking how is a consensual echo propagated through the AMT. To build a consensus, the

[7] A technical term in WordNet's parlance for the description of a synset.

workers develop a sense of the speed required by the apparatus. This speed as I have said is induced by the remuneration, but it is also induced by many more elements. For instance, AMT stabilises the candidate images as thumbnails detached from their original context. They are search engine results and as such the search engine already conveys a sense of consensus. If many candidate images look similar, it suggests that a consensus exists over them. The Turkers do not validate mechanically the dominant representations they are given by the interface, but the regularities that spring out from the grid of thumbnails function as a cue. It is an accelerator. To choose against the coherence that emerges from the results requires more work and time. It requires looking at the candidates that do not stand out. Another important cue comes from the jobs being accepted or refused by the requester. The reasons why a job is accepted (understand, remunerated) or refused (the worker has no recourse against the requester) are extremely rarely given by the requester. It is, therefore, not always possible to correlate a decision made by the worker and a rejection from the requester. But as they have significant economic consequences for the annotator, she evaluates her work based on her interpretation of the requester's decision. Therefore, the Turkers never rely on any explicit guidance, but they follow an echo, where different forms of feedback and cues resonate with each other. Workers need to learn how to listen to the apparatus more than they have to read the labels and the definitions of the synsets.

The Parisian or the reformer cases make clear that the composition of the synsets cannot be explained by a process in which workers carefully read the gloss and select the candidate images accordingly. Focusing on a pivotal moment of an "isolated who" making a decision takes us only so far. The Turkers hover over a visual configuration emanating from the interface. Rushing through the pages, they see an overwhelming presence of the shining blondness of a familiar icon whose name matches the synset's label. The Turkers do not necessarily believe that Paris Hilton is a resident of, or born in, Paris. Their decision is that there are enough responding echoes in the apparatus to validate the selection without having to spend time further verifying. The object of the consensus is not the fact that Hilton is a Parisian, but that such an approximation will not isolate them. A course of recognition traverses the apparatus. To recognise is to be recognised. To be recognised as a Turker, one doesn't need to recognise things that are factually right but to recognise the grade of approximation that is expected. The Turker's competence is multi-scalar, not just a semantic affair. As a Turker wrote on a forum, it is to have a sense of what is "enough to get away with". To get away is indeed the right term as a Turker must always have an eye on the previous hit and another on the next. Speed is conductive to consensus.

6 Arbitrating photographic alignments

Following the production of consensus in the annotation environment, we gained insights into the nature of the workers' transepistemic contribution. They are accomplishing speed and accuracy rather than emitting explicit judgements. They are constantly engaged in probing the plasticity of the consensus rather than producing discrete statements about facts. Having established that, I need to relate the workers' contribution to the alignment they are part of. I have already stated, following Crimp that the condition for the diversity of the data set is that the objects are represented in various conditions but that the medium which represents them is uniform and transparent. Yet as we have seen the workers are involved in more than classifying concepts represented differently through a seamless medium. The heterogeneity of the candidate images doesn't only reflect a disparity of styles and contexts. It also reflects the heterogeneity of larger apparatuses through which photographs are stabilised and set in motion. To understand the full consequence of this, we need to come back to the notion of alignment and the role played by the search engine in the data set's acquisition pipeline. The search engine hides the relations and context in which a photograph lives online. It excises the photograph from its alignment and, therefore, contributes to the appearance of the photographic object being independent from the apparatus through which it is made visible. This extraction is also in part the validation of the authority of an apparatus over a query term. In many synsets there are large amounts of photographs from the same source (e.g., a Flickr album, a blog or website). The distribution of the photographs in a synset reflect the effectiveness of the strategies of the various platforms competing for the visibility of their contents. A platform like Flickr very much anticipates the requirements of a search engine and offers it as a structured set of objects (file, metadata) optimised for machine readability. By securing and reinforcing the internal connection of its content (Flickr photos are related to each other by various mechanisms like links, inclusions in groups or tags), the platform also anticipates the search engine's ranking criteria (i.e., the more links the higher visibility). The platform is "search engine optimised" and the users of the platforms are recruited in this effort. The search engine and the source platform are not clearly delineated entities between which independent JPEG files are in transit. Rather the platform very much interiorises the search engine in many ways. This explains that various categories are "taken over" by content popular on a sharing platform as the platform is optimised to do so.

In this competition, platforms have an advantage as they are built from the start to capitalise on the circulation of

Illustration 3 synset n04070003, Reformer

their assets. But this advantage is only partial. For certain categories, other apparatuses beat the platforms at their game. The monopoly then comes from other sources. The synset n09633969 "wrongdoer, offender", for example, is essentially constituted from mugshots published by US states or county websites. These administrations operate their own apparatuses of capture, identification and dissemination that dramatically contrasts with those of amateur photography. These apparatuses inscribe their own regularities in terms of pose, colours and point of views. They also inscribe other regularities in terms of race and gender by excluding images of white collar (or female) criminals and focusing on people of colour. As a consequence, a synset like wrongdoer, constituted of mugshots does not merely validate photos that are imaging a concept, it perpetuates the politics of shaming inherent to state department's websites.

The data set is not a merely a collection of photographs, it is an arbitration between different forms of alignment: Flickr's amateur snaps, police portraits, submarine photographs are re-aligned as items of a training set. It doesn't make sense to talk about this process of selection as if a group of distinct individuals were choosing discrete photographs in a void, but as the data set feeding on different photographic alignments, branching itself into various apparatuses and offering them different shares of the data set's space, and delegating various levels of decisions to the search engine algorithm and the AMT workers.

Having followed this process of realignment, it is clear that using photography as a tool to homogenise the visual world leads to a paradox. The paradox is that to resolve a photograph as data, all the heterogeneity that pertains to the medium, its apparatus and its circulation needs to be repressed. The photograph needs to be *made* a transparent vehicle. And to make it transparent, computer scientists need to engage in the production of their own apparatuses and produce their own alignments. This long chain of translations and ruptures from search queries to AMT to data set is in itself an alignment and also conditioned to the existence and performance of apparatuses. The paradox, therefore,

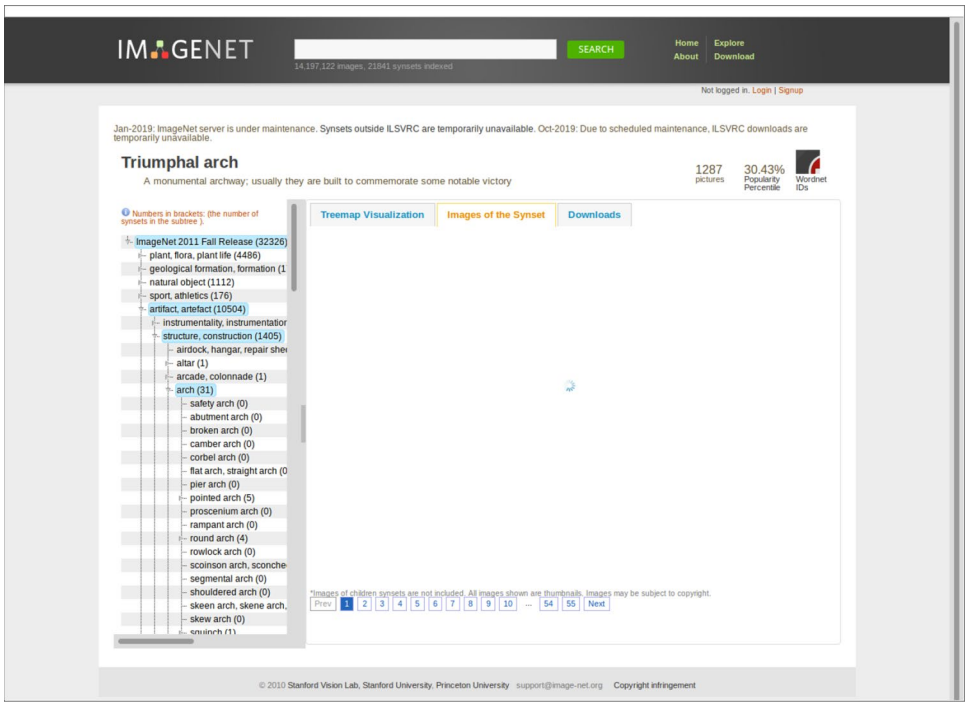


Illustration 4 A screen capture of the Triumphal Arc synset on imagenet.org

lies in this: it is an alignment that needs to negate itself, that needs to remove itself to construct photography as a transparent leveller of data. The workers are, therefore, engaged in a double process of elision. Their own contribution is unacknowledged, as they are considered to be merely mechanically responding to hits. Their contribution consists of erasing the traces of the apparatuses involved in the resolution of online photographs into data set items. The i*nvisibilisation* of the workers goes hand in hand with the resolution of photography as a transparent leveller of data.

7 Conclusion

At this point, I have made the case that engaging with the photographic elaboration of computer vision requires the acknowledgement of the importance of the workers epistemic contribution and the photographic alignment traversing CV. To conclude this paper, I will show how this notion of photographic elaboration can help question the terms by which current controversies in CV are being framed and the responses that have ensued (Illustration 4).

Despite its immense academic success, its inclusion in many pieces of software, and its celebration as a benchmark for AI algorithms, ImageNet, today, is undergoing a grave crisis. For those who consulted its website during the last years, ImageNet's online presence has been gradually disappearing. Its disappearance is due in larger part to mounting controversies both from the public (for instance, see *ImageNet Roulette*, Crawford and Paglen 2019) and from inside the tech community (i.e., Dulhanty and Wong 2019; Shankar et al. 2017; Recht et al 2019). The criticism mainly targets the data set's cultural, racial and gender bias. In response, the data set creators and a larger team of researchers are engaged in a process of revision of the data set as an attempt to remedy to the various shortcomings expressed in and out the CV community.

Current reparative efforts are concentrated on the mainstreaming of the training data, the objective being to make ImageNet fairer and updating its "acquisition pipeline". The researchers' response to the problem, as expected, essentially concentrates on the question of representation. Problematic categories are annotated to attach properties such as gender, age or skin colour to images. To give fairer representations means, for instance, rebalancing the roles portrayed in the photographs according to gender or racial criteria, and to produce a more suitable distribution of these roles in a synset. Leaving aside the questions of how criteria such as gender or skin colour are assessed, or what a suitable distribution of photographs based on such criteria might mean, it is clear from the outset that the response once again concentrates on the depicted objects to evaluate how well they image a concept. As a response to external pressure, the representationalism of the data set makers increases. The "People" subtree is annotated to evaluate the imageability of its concepts. All the concepts with a low imageability grade are removed. The solutions designed by the research team keep on reducing the number of selected elements in the data set, and potentially problematic categories are discarded. Data demining is in order. At the end of the process, the

researchers estimate that out of the 2832 subcategories of the "people" subtree, only 139 will remain.

But as we have seen, the problem runs deeper than representation. A photograph is not just a constructed representation, it is also very much defined by its circulation, the currencies in which it is accepted, how it comes into series. It acquires meaning and function through use. It requires a reflection on the conditions in which a photograph is temporarily resolved, to the different elements it aligns with. It requires a renewed attention to the relation with the apparatuses and the engineer's own apparatus of capture, his/her own alignments. Therefore, a response that addresses a criticism of bias only narrowly defines the problem. If the problem is understood solely as a question of right or wrong data (more balanced and fairer), it will leave unquestioned how photography is made to be data in the first place. It remains blind to the performative role of the methods, the labour, the agents, the scales, the apparatuses and the alignments it relies upon.

If we acknowledge that photographs are not representations independent from their alignments, what is the consequence for data set makers? If the process of production of the data set doesn't take into account the whole chain of aligned elements it rests upon, the nature of its own apparatus, it will keep treating visual data as a mere collection of elements that can be replaced by safer ones or entirely avoid risky categories. But this improvement will then necessarily lead to further impoverishment of the data. As an apparatus in denial of its own performativity will only neutralise further the relational nature of the photograph. Engineers apply the GIGO maxim, "garbage in, garbage out", and imagine getting rid of the garbage is the path forward. Accordingly, a criticism based on bias will lead to an increased cleaning of the data set. But image *is* garbage. It is entangled in conflicting regimes of representations, divergent apparatuses, disruptive alignments and irreconcilable practices. If those are cleaned up, there is no image left. There will be perhaps less risk for algorithms to be contaminated by obscene or offensive imagery if the data set keeps on shrinking. But this entails a bigger risk, the risk for machine learning to not learn anything worth knowing. A data set of perfect representations is an empty one. The most striking images ever produced by computer vision are not deep, either fakes or dreams, they are flat and white and on their centre, an icon suggests the page is loading without ever displaying any content. They are the millions of empty pages of the ImageNet website. These images of the data set's ruins are the products of its inability to engage with the labour it relies on and the apparatus it is entangled with.

To introduce this text, I referred to Crimp's take on the musée imaginaire. I stated that the ruins Crimp refers to were not solely the ruins of extracted elements dying in the confines of a mausoleum. They were the ruins of a project that relied on a particular change of scale. Malraux's project, Crimp stated, was an hyperbole representing the transformation of the museum. The "real" museum is seen as limited in what it can acquire and keep within its walls. Photography makes it possible to enlarge the scope of what one sees and how one sees. The museum as imagined by Malraux aims to produce knowledge by comparison. Works can be shown side by side, monuments can be scaled down and details can be enlarged. Before, one would compare a painting with a memory, now one can compare two objects through their photographs. The problem Crimp addresses is that photography is enrolled in the process of articulating the relations between a potentially unlimited amount of works on the condition that its role should be limited to one of revealing. Therefore, the process of comparison of otherwise heterogeneous elements is only meaningful, because the process of extraction and homogenisation is made invisible. But as Crimp ironically remarks, when photography enters the museum as an art object, heterogeneity returns as: "Even photography cannot hypostatize style from a photograph" (Crimp 1980, p. 53).

Crimp's lesson is that we must extend our concern for the objects to the larger project, to the investment in a certain scale obtained through photography and the putative knowledge it purports to generate.

Instead of lamenting the presence of dead objects in a mausoleum, one should question the ruins of an ideal plane of comparison. To translate this in the context of CV, instead of taking off every element that could lead to controversy, the data set makers should instead reckon with their own apparatus of annotation and the labour it conceals. Instead of denying the collaboration, they should address the transepistemic dimension of the work carried out by the annotators and engage with it. This is hard work, because it questions the discipline, where it hurts most: at the level of its economy. Recognising the work carried out by the Turkers obviously opens the question of the financial revalorisation of their labour. Furthermore, computer scientists should consider afresh their understanding of photography and do away with the entrenched representationalism of the discipline. This is hard word again, because it rejects the notion of an ideal plane on which all images can be made data. Combined together these two observations make clear that the work to be done is much more ambitious than bias-proofing sensitive categories. It confirms Li's insight: data sets are indeed tools to resolve the scale of computer vision's problem. However, it suggests that resolving the scale is a project that needs a fresh start.

Acknowledgements This article is based on a research conducted thanks to the support of London South Bank University and The Photographers' Gallery. It has benefited from the attentive comments of Gaia Tedone, Katrina Sluis, and Ruben Van de Ven. It draws on a reflection shared with Laurence Rassel, Geoff Cox, Andrew Dewdney

and my colleagues at the Centre for the Study of the Networked Image, as well as the Institute of Computational Vandalism and the Constant collective.

References

Beheshti S-M-R, Tabebordbar A, Benatallah B, Nouri R (2016) Data curation APIs. CoRR abs/1612.03277

Brownlee J (2019) A gentle introduction to computer vision. https://machinelearningmastery.com/what-is-computer-vision/. Accessed 25 Mar 2020

Crawford K, Paglen T (2019) Excavating AI: The Politics of Images in Machine Learning Training Sets. https://www.excavating.ai/. Accessed 19 Mar 2020

Crimp D (1980) On the museum's ruins. October 13:41–57. https://doi.org/10.2307/3397701

Daston L, Galison P (2007) Objectivity. Zone Books, New York

Deng J, Berg A, Li K, Li F-F (2010) What does classifying more than 10,000 image categories tell us? In: ECCV. pp 71–84

Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: In CVPR

Dulhanty C, Wong A (2019) Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets. CoRR abs/1905.01347

Fei-Fei L (2010) ImageNet, crowdsourcing, benchmarking & other cool things. In: CMU VASC Seminar

Fuller M (2005) Media Ecologies: Materialist Energies in Art and Technoculture. The MIT Press

Fullscale (2019) Machine learning in Computer Vision. https://fullscale.io/machine-learning-computer-vision/. Accessed 25 Mar 2020

GoogleTechTalks (2011) Large-scale Image Classification: ImageNet and ObjectBank. https://www.youtube.com/watch?v=qdDHp29QVdw. Accessed 28 Feb 2019

Hara K, Adams A, Milland K, et al (2018) A data-driven analysis of workers' earnings on amazon mechanical turk. In: proceedings of the 2018 CHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 449:1–449:14

Hata K, Krishna R, Fei-Fei L, Bernstein MS (2017) A glimpse far into the future: understanding long-term crowd worker quality. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. ACM, New York, NY, USA, pp 889–901

Irani LC (2015) The cultural work of microwork. New Media Soc 17:720–739. https://doi.org/10.1177/1461444813511926

Irani LC, Silberman MS (2013) Turkopticon: interrupting worker invisibility in amazon mechanical turk. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 611–620

Karpathy A (2014) What I learned from competing against a ConvNet on ImageNet. https://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/. Accessed 10 Jul 2019

Knorr-Cetina KD, Malkay M (1983) Science observed: perspectives on the social study of science. Sage, London

Malraux A (1965) Le muséeimaginaire. Gallimard, Paris

Marr D (1982) Vision. A computational investigation into the human representation and processing of visual information. MIT Press, Cambridge, MA

Mol A (2002) The body multiple: ontology in medical practice. Duke University Press, Duke

Paivio A (1986) Mental representations: a dual coding approach. Oxford University Press, New York

Papadopoulos DP, Uijlings JRR, Keller F, Ferrari V (2016) We don't need no bounding-boxes: training object class detectors using only human verification. CoRR abs/1602.0

Quach K (2019) Inside the 1TB ImageNet data set used to train the world's AI: Naked kids, drunken frat parties, porno stars, and more. https://www.theregister.co.uk/2019/10/23/ai_dataset_imagenet_consent/. Accessed 25 Mar 2020

Recht B, Roelofs R, Schmidt L, Shankar V (2019) Do ImageNet Classifiers Generalize to ImageNet? https://arxiv.org/pdf/1902.10811.pdf. Accessed 10 Jul 2019

Rosenberg C (2013) Improving Photo Search: A Step Across the Semantic Gap. https://ai.googleblog.com/2013/06/improving-photo-search-step-across.html. Accessed 27 Feb 2019

Shankar S, Halpern Y, Breck E, et al (2017) No classification without representation: assessing geodiversity issues in open data sets for the developing world

Simpson Center for the Humanities UW (2017) Lorraine Daston on Algorithms Before Computers. https://www.youtube.com/watch?v=pqoSMWnWTwA. Accessed 25 Mar 2020

Smith C (2019) 20 Interesting flickr facts and stats 2019 by the Numbers. https://expandedramblings.com/index.php/flickr-stats/. Accessed 10 Jul 2019

Torralba A, Fergus R, Freeman WT (2008) 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Trans Pattern Anal Mach Intell 30(11):1958–1970. https://doi.org/10.1109/TPAMI.2008.128

Vijayanarasimhan S, Grauman K (2009) What's it going to cost you? Predicting effort vs. informativeness for multi-label image annotations. In: 2009 IEEE conference on computer vision and pattern recognition. pp 2262–2269

Wikipedia Computer Vision. https://en.wikipedia.org/wiki/Computer_vision#Related_fields. Accessed 25 Mar 2020

Yang K, Qinami K, Fei-Fei L, et al (2019) Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy