AI & SOCIETY (2021) 36:1153–1165
https://doi.org/10.1007/s00146-020-01091-y

ORIGINAL ARTICLE

On machine vision and photographic imagination

Daniel Chávez Heras[1] · **Tobias Blanke**[2]

Received: 1 May 2020 / Accepted: 14 October 2020 / Published online: 17 November 2020

Abstract
In this article we introduce the concept of *implied optical perspective* in deep learning computer vision systems. Taking the BBC's experimental television programme "Made by Machine: When AI met the Archive" (2018) as a case study, we trace a conceptual and material link between the system used to automatically "watch" the television archive and a specific type of photographic practice. From a computational aesthetics perspective, we show how deep learning machine vision relies on photography, its technical regimes and epistemic advantages, and we propose a novel way to identify the *latent camera* through which the BBC archive was seen by machine.

Keywords Computational aesthetics · Philosophy of photography · AI television · Computer vision · Deep learning · Dataset archaeology

1 Introduction

Is that a person or a reflection? A man or a woman? Is the woman holding a mobile phone, or is it, rather, the statue of an ancient Egyptian king? Is the man wearing a shirt, or is it an elephant? Or a stuffed animal holding a banana…? (Figs. 1, 2, 3).

These are some of the mislabellings produced when a small team of technologists and researchers set a computer vision system to "watch" thousands of hours of British television for the project "Made by Machine: When AI met the Archive"(MbM), whose outputs were eventually packaged and broadcast on BBC Four as an experimental "AI TV" programme in 2018.[1]

In line with the public purposes of the British broadcaster (BBC 2018), one of the main goals of the programme was to show to a wider audience some of the possibilities and limitations of AI, and in particular deep learning approaches that underlie many contemporary computer vision systems. From a research perspective, the project was also designed as prompt to explore just how exactly computers are said to be "seeing". What type of knowledge is produced by computer vision and how does it inform the ways we understand and give currency to audio–visual media more generally?

In related work, such questions have generally been approached by focussing on training datasets and how they are assembled as well as how the resulting AI systems represent or fail to represent different sectors of society. Exemplary of this approach are the works of Kate Crawford and Adam Harvey:

> "Training sets, then, are the foundation on which contemporary machine-learning systems are built. They are central to how AI systems recognize and interpret the world. These datasets shape the epistemic boundaries governing how AI systems operate, and thus are an essential part of understanding socially significant questions about AI." (Crawford and Paglen 2019)

> "A photo is no longer just a photo when it can also be surveillance training data, and datasets can no longer be separated from the development of software when software is now built with data." (Harvey and LaPlace 2019)

Research using this approach has shown datasets to be deeply problematic, both in their politics of assemblage and of representation. The experience in MbM was no exception here. A team from BBC R&D implemented an automated dense captioning system pre-trained on the Visual

✉ Daniel Chávez Heras
daniel.chavez@kcl.ac.uk

[1] Department of Digital Humanities, King's College London, London, UK

[2] University of Amsterdam, Amsterdam, The Netherlands

[1] https://www.bbc.co.uk/programmes/b0bhwk3p

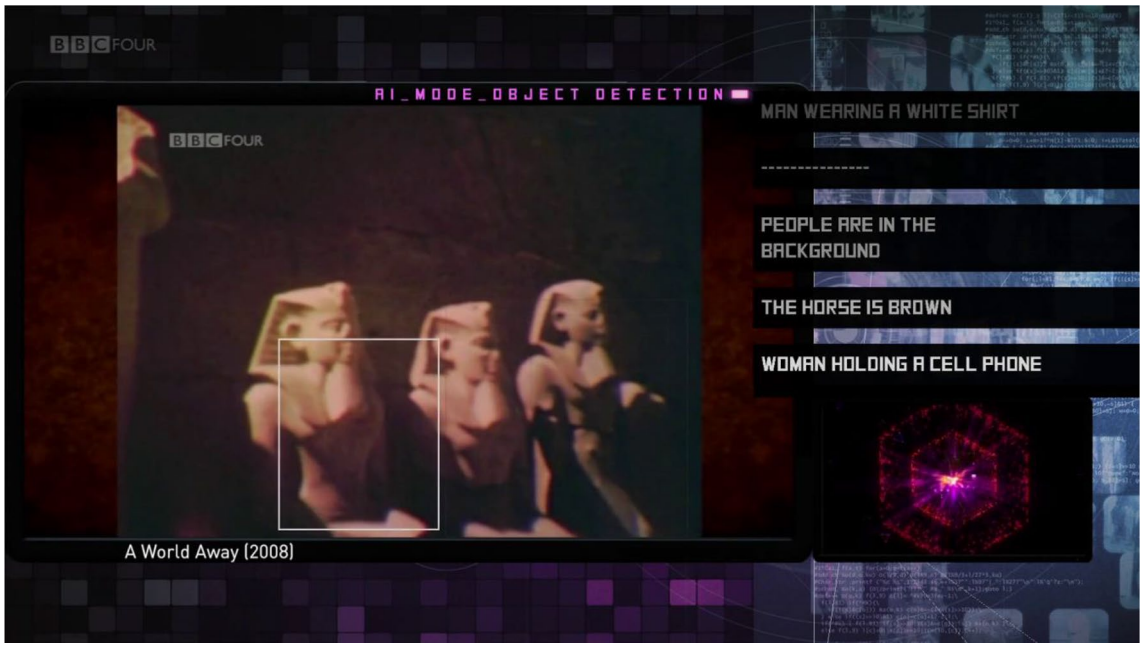

Fig. 1 MbM still frame with predicted label 'woman holding a cell phone'

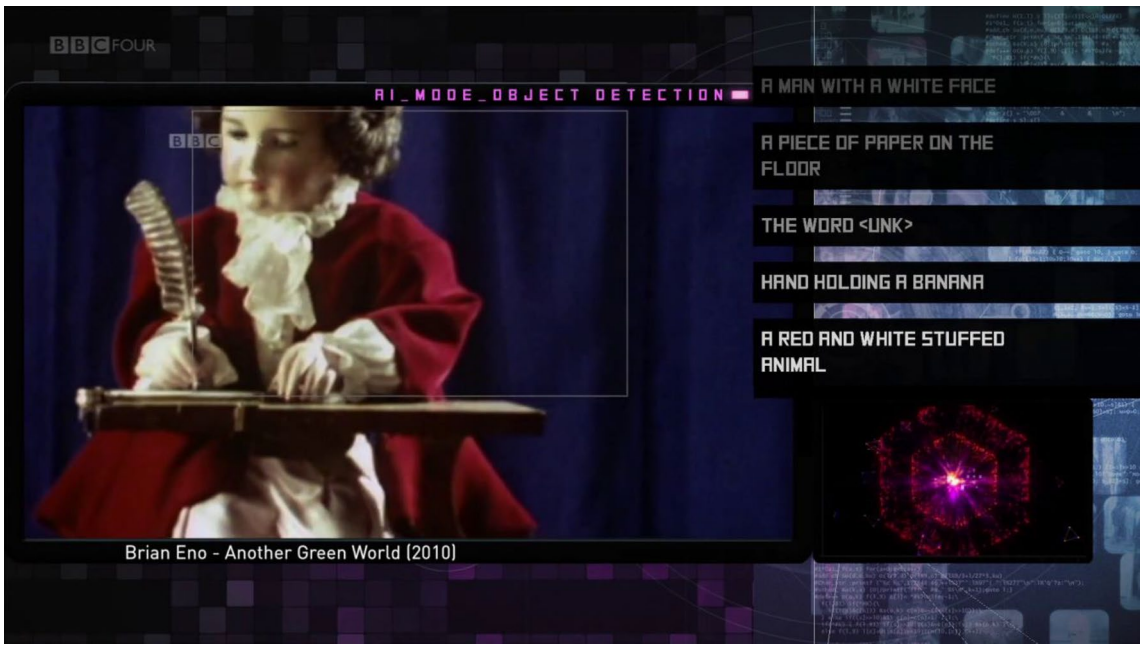

Fig. 2 MbM still frame with predicted label 'red and white stuffed animal'

Genome dataset[2]; comprised of little over 108,000 images downloaded from Flickr and annotated by 33,000 Amazon Mechanical Turk workers, 61% of whom were under 35 and 93% from the USA (Krishna et al. 2017, 43). This captioning system implementation was used to automatically annotate several hundred hours of television from the BBC archive, and these annotations used as metadata in an associational engine that concatenated new sequences of related television clips (Cowlishaw 2018; Chávez Heras et al. 2019).

[2] https://visualgenome.org/

Fig. 3 MbM still frame with predicted label 'reflection of a mirror'

One of the results immediately observed in these machine-generated clips was the system's propensity to identify men in shirts. Although imbalance in gender representation is a known issue in television, including the BBC (Cumberbatch et al. 2018), this effect was so pronounced that it prompted a closer inspection of the training dataset. "white" is the most common attribute, "man" most common object (with twice as many instances as "woman") with "shirt" also among the top ten (Krishna et al. 2017, 50–53, 63). Biases such as these have been consistently found in this and other large image training sets used in deep learning computer vision, the localisation and disaggregation of such biases being one of the main goals of Crawford's proposed archaeology of datasets (Crawford and Paglen 2019).

However, by focusing on labelling images as the main layer of human intervention, i.e. the principal source of data, bias and error, such an archaeology very often overlooks other significant areas of human subjectivity encoded in these systems, namely the nature of the images themselves. These labels are produced not over direct observations of the world, but over photographic images of it; images that are technically and socially mediated in powerful ways that create and sustain specific regimes of visibility and with which we hold a complex relationship before they become digital artefacts.

Following this line of thought, we set out to look at computer vision through the (figurative and literal) lens of photography. Through a technical genealogy of photographic lenses, the types of images they afford and the social functions given to these images, we show how AI is materially and conceptually connected to optical regimes of visibility. Based on this connection, in the last section we assemble a dataset with which to analyse photographic practice, and then use it to train a bespoke focal length classifier as a proof-of-concept for a system designed to investigate the optical perspectives implied in MbM.

2 Epistemic discontents

An often unexamined fact about of deep-learning computer vision is that the millions of pictures it algorithmically mobilises are, for the most part, photographs. Perhaps one of the most revealing accounts about how this choice came to be seen as obvious comes from Fei-Fei Li, one of the creators of *ImageNet*[3] and a key figure in the shift towards deep learning over the last decade. Li was asked recently about the choice of using photographs for *ImageNet* during an event celebrating the tenth anniversary of the dataset: "That's a great question." —she replied— "We didn't really stop to think much about it (…). I suppose we wanted as a realistic representation of the world as possible" (Li 2019).

Li is not alone in her assumption about realism and photography. A widely shared intuition about photographic images is that "the camera does not lie" or that in any case it lies less than other methods of depiction. In an often-cited passage of his influential *Ontology of the Photographic*

[3] https://www.image-net.org/

Image, André Bazin (1960[1958]) wrote that the invention of photography and cinema "satisfy, once and for all and in its very essence, our obsession with realism. No matter how skillful the painter, his work was always in fee to an inescapable subjectivity. The fact that a human hand intervened cast a shadow of doubt over the image." (7). What he meant exactly by "realism" has been the subject of much debate, since but this view of photographs as trusted visual renderings of the world due to their alleged automatic mode of production has proved remarkably enduring.

Li's response earnestly voices just such a view, where unlike for example drawings or paintings, which are inextricably bound to the mental states and technical abilities of their authors (as well as the embodied command of these states and abilities), photographs appear to be pictures produced through mind-independent processes. They capture whatever is in front of the camera regardless of what the photographer believes about what is in front of the camera, i.e. cameras can only show what is there to be seen. In philosophy, this idea of mind-independence mechanical process has served to explain the epistemic advantage of photographs over other types of images (Cavell 1979; Cohen and Meskin 2004; Walden 2005; Abell 2010).

One formulation of this argument proposed by Gregory Currie (1999) is that we treat photographs as *traces* as opposed to *testimonies*. The former are counter-factually dependant on nature, like a footprint, in a way that the latter are not, like the tale of how someone once took a step in the mud. For the footprint to be any different, Currie would argue, the sole of their shoe would have had to differ accordingly, while a description of the step taken, however, rich or detailed, necessarily implicates the intentions of the describer and belongs, therefore, to a different epistemic register altogether. According to this view, the social credibility lent to photographs makes them more akin to *light detections* captured as a result of a mind-independent mechanical process, while paintings and other pictures made by hand tend to be seen as *someone's depiction* of a scene, this is, as the result of an embodied cognitive and creative process.

This is the dominant logic that underwrites computer vision too, at least in its current form and insofar as it is powered by photographic images, from which it inherits, exploits and amplifies their epistemic advantage founded on this mind-independent conjecture about the photographic process. When millions of photographs are aggregated into large datasets and used to train machine learning systems, the representational powers of photography and computation compound, to the point, where predicted labels are also seen as traces, face *detection* and not face *depiction*. The predictive tokens produced by computer vision are thereby presented as the counter-factually-dependent *and* mind-independent detections of something or someone. This person or this object was seen automatically and, therefore, *had* to be there to be seen.

Recently, however, this view has been put under mounting pressure by the so-called new theory of photography, whose proponents argue for an expanded view of the photographic event as a multi-staged process of which only some parts can be said to occur automatically (Maynard 1997; Phillips 2009; Lopes 2016; Costello 2017). Costello (2017), for example, identifies a contradiction in ascribing epistemic value to photographs on the basis of their supposed mind-independence processes while simultaneously characterising these processes as automatic. A process cannot be both natural and automatic, he argues, without separating humans from nature (42). Automatic processes are causally-dependant but not spontaneous according to Costello. For a process to be called automatic it should be possible to specify it in terms of the labour that is being delegated to a mechanism, one that serves human ends and, therefore, necessarily involves human minds. Costello asks:

> "just what is it exactly that is supposed to 'happen by itself'? […] In photography, almost everything that expresses comparable choices [to painting] happens *off* the support—the choice of lens, distance, lighting, moment of exposure, point of view, etc.[…] This can give those who have no idea, where to look the impression that the photographer has done very little, or that the mechanism is responsible. But this is plain ignorance. The fact that so many of the acts take place prior to the image appearing 'all at once' does not negate the photographer's responsibility for what then appears. Merely noting depth of field markers in an image already tell us much about what the photographer was after. One needs to be a competent judge in photography as in any other domain; and this requires a basic grasp of the internal relations between focus, depth of field, and exposure that most Orthodox theorists fail to evince." (45) [italics in original].

From this perspective, photographs are seen as faithful visual representations by virtue of their mechanisms no less than by the ways in which such mechanisms are controlled and regulated by photographers. Following Costello, we need to also consider that our intuition of what is a "realistic" image rests in the case of photography as much on *what* it shows than on *how* it shows it, which is to say that the photographic image is granted its privileged epistemic position in society by adding, not subtracting layers of subjectivity; not "without the creative intervention of man"(7), as Bazin would have it, but precisely because of it.

If the depicted is indeed inextricable from the process of depiction, we must then ask why should we not care about this process when it comes to computer vision?

3 The glass computer

One of the reasons photographic cameras appear to record the world automatically is that many of the calculations needed to render space visible are pre-programmed in the photographic devices themselves, most significantly in photographic lenses. When photographers *pull* an image into focus by adjusting the focus ring, the lens is doing some of the heavy lifting in terms of the calculations necessary to harness light convergence and render a slice of space visible in a specific manner. This is not to say the lens itself thinks, but rather that thought has been put into the lens, quite literally crystallised in its design, and that the photographer is able to interface with it through the camera controls.

From a cognitive standpoint, like the Sumerian abacus or the medieval volvelle, photographic lenses can be thought of as a type of analogue computer: a system that allows its user to actively externalise memory to a programmable calculating object and in this way distribute the cognitive load required to perform a specific task. Configured in this manner, user and object enter into an interaction feedback loop, creating "a coupled system that can be seen as a cognitive system in its own right" (Clark and Chalmers 1998). That photographers need not perform optic calculations to mobilise their effects is one of the most salient affordances of photography as a technology, and connects it to computer vision in their shared overarching project to automate visual labour through the computation of space (Pasquinelli 2019), with the obvious difference that in the case of photographic lenses such computation is analogue.

If we have not traditionally thought of photographic lenses as machinery with which to calculate[4] is perhaps because their inputs and outputs are presented as images and not numbers or letters. We do not know, for example, whether an image is in focus if presented as a matrix of pixel values (or a tensor); we need to see the results as an image "all at once" to evaluate it. However, the intermediate steps of interaction involved in producing photographic images are in fact heavily mediated by numerical parameters and standardised metrics: focal length, exposure, aperture and ISO. These are all given as numbers that describe the internal relations of the photographic mechanism, and from this point of view a key aspect of photographic practice as an imaging technology is to understand, control and harness different permutations of these relations for a variety of lenses. An equivalent process in machine learning would be the understanding and control of hyper-parameters.

Take focal length as an example. Today, it is widely used as a standard measure for lens classification, since it correlates with the size of the image plane[5] and the aperture[6] of the camera to define, among other things, the field of view, i.e. how much of a given scene fits into the frame, and the depth of field, i.e. how much of it is in focus at any given time. For a full-frame format,[7] a wide angle lens (e.g. 28 mm) will cover a wider field of view and have a deeper focus range, while a telephoto (e.g. 300 mm) will magnify to a narrower area and have a shallower focus range. In between we find a 50 mm, often called a "normal" lens.

Over time, the effects produced by different focal distances get thematised and are attached to specific social narratives. Long telephotos tend to be used in sports and nature photography, where subjects are often moving at a distance and backgrounds can be out of focus. Wide lenses, on the other hand, privilege field of view and focused scenes instead of magnification. Depicted through a different lens the same subject can be made to look *in fraganti* in a leaked mobile phone picture (28 mm) or like a model fit for the cover of a fashion magazine (175 mm) (Wieczorek 2019).

Different lenses contribute in this way to our understanding of what pictures are *about*. Our argument is that it is precisely this "aboutness" of vision that we feel to be conspicuously absent or compromised in the tokens of prediction produced by systems like the one used to machine-see the BBC television archive, a type of computer vision which only points to what images are *of*.[8] Drawing from this distinction, we can clearly see how lens aesthetics are not incidental to photography but rather a fundamental dimension of its epistemic advantage insofar as they enable distinct relations between the see-able and the know-able; between knowledge and the appearance of knowledge. That images are seen to be *about* something inasmuch as *of* something suggests that we put our faith in photography not because it offers undistorted images of the world, but because we

[4] However there are a lot of optical calculations crystallised in photographic lenses. One notable example that links optics with computing is how in 1840 Joseph Petzval, an Austrian mathematician, employed several human computers to aid in the design a new four-element lens capable of under-one-minute exposures: the famous *Petzval Portrait* (See: Peres 2007, 159).

[5] Usually given in millimetres as the diagonal measure of a rectangular projection surface or screen onto which an image is formed when reflected light is projected through a lens.

[6] Known as *f*-stop or, somewhat confusingly, *f*-number (N), calculated using the formula: N = f/D (where f is focal length and D the diameter of the iris or pupil that allows light into the lens.

[7] 35 mm (36 mm × 24 mm frame) film negative or equivalent digital sensor. Many digital cameras have smaller sensors, thereby modifying (cropping) the field of view of lenses. The smaller the sensor is in relation to a full-frame the larger its crop factor. Conversely, by knowing the crop factor of a given sensor, one can estimate a lens' full-frame equivalent focal length. Mobile phones, for example, have much smaller sensors and lenses, an iPhone X has a 4 mm lens, 28 mm equivalent in a full frame camera.

[8] For a discussion on this distinction see (Maynard 1997).

believe that photographic distortions are meaningful. Computer vision, we argue, gains it powers by treating photographs not as detections of the world, but as measurements of these beliefs, and in doing so it assumes an implied optical perspective.

4 A machine made of images

Let us now return to MbM and ask what optical perspective is implied in it. What lens or lenses are encoded in the computational gaze we set upon the BBC Television archive?

We know the Visual Genome uses images originally sourced from *Flickr* (Krishna et al. 2017, 47) and that the photo platform hosts many of its images along with their EXIF data,[9] which is an international metadata standard for digital images and sound that includes tags for camera settings and lens information.

EXIF is far from perfect. Its metadata structure is borrowed from TIFF files and is now over 30 years. A notable drawback to working with this type of data is, therefore, its inconsistency, given the quick pace at which digital cameras changed over the last decades and the many differences in how they used the standard over time, even among cameras from the same manufacturer. What is more, some manufacturers like Nikon use custom format fields not common to any other brand and encrypt the metadata contained in them. This makes it very difficult to extract, disaggregate and process EXIF.[10] Finally, this type of metadata is not usually available for not-born-digital photographs, i.e. taken with analogue cameras or images that were scanned.[11]

These caveats notwithstanding, EXIF is still the most widely used metadata standard for photography and as such a key resource to research the equipment and technical practice that underlies the creation of photographic images in the digital age. And precisely because of its longevity and pervasive use, it is one of the few ways to trace a technical lineage from lenses to computer vision. It is quite possibly the only, where such a lineage can be done at a larger scale, given the size of the collections of images used in deep learning.

Table 1 Example of Exif tags extracted

Tag	Value
Camera manufacturer	Canon
Camera model	Canon EOS 7D
Exposure	1/1600
Aperture (F-number)	2.8
Focal length	145.0 mm
Lens info	0EF70-200 mm f/2.8L IS II USM

We extracted EXIF metadata from all the images whose Flickr IDs matched the ones present in the Visual Genome. The metadata standard is comprised of over twenty thousand tags, but we only selected tags that were general enough so as to be reported by most cameras. Within these, we only focused on the ones containing data about the parameters over which photographers tend to have more choice and control, namely their choice of camera and lens, as well as the aperture, exposure and focal length settings. Table 1 shows a list of the tags that were queried and an example of the values extracted. Table 2 shows an overview of the extraction results.

The extraction process yielded a relatively dense distribution, with over 83% of accessible images returning metadata in at least one of the five of the queried tags. The one exception was < Lens info >, for which only 10% of accessible images returned values. In light of this, we decided to consolidate data for all tags except < Lens info >, which was kept separately for later analysis. We also parsed over apertures and focal lengths to bin them into categories: twelve bins corresponding to full *f*-stops for apertures—from *f*1 to *f*45,[12] and seven focal distance bins corresponding to a commonly used classification[13]:

[9] Developed in 1998 by the Japan Electronic Industries Development Association (JEIDA), eventually absorbed by the Japan Electronics IT industries association (JEITA) and the Camera & Imaging products association (CIPA). (See: JEITA Standards).

[10] https://exiftool.org/TagNames/Nikon.html#LensData01

[11] A scanned photo, for example, will sometimes include EXIF data from the scanner, but obviously no lens or other information about the camera with which was originally taken. It is possible, however, to manually write or re-write EXIFtags of a digitised photograph (or nearly any other digital image for that matter). For the most authoritative source on working with EXIF see Phil Harvey's Exif Tool: https://exiftool.org/

[12] There are wider and narrower apertures, for example *f*/0.95 of the Shenyang Zhongyi Mitakon Speedmaster 50 mm. However these are very rarely encountered and were not observed in our dataset.

[13] These categories are not policed or enforced by any particular institution, as the boundaries are seen as irrelevant in most areas of photographic practice. They are, rather, more of a tacit agreement among photographers, lens and camera manufacturers. Of the tags queried, focal distance was the most challenging because, as we noted earlier, these values are relative to the size of the sensor. The standard "full frame" sensor was adopted as an equivalent of 35 mm film stock, but as digital cameras shrunk in size so too did their sensors. The effect is particularly stark in mobile phones, whose sensors are particularly small, so in order to compare their focal length to that of larger cameras one needs to multiply their reported Exif value by a crop factor so as to obtain a 35 mm equivalent. This crop factor is different across models and manufacturers, for example many Apple iPhone models have a sensor crop factor of 7.6, if the focal length in their Exif metadata is 4.3, the 35 mm equivalent is a little over 30 mm. If one were to accurately measure focal length one

Table 2 Overview of the extraction results

Category	Count	Percentage
Total number of IDs processed	103,077	100.00%
Unavailable URL request (500 error)	18,521	18.00%
Available image but with no data in the queried tags	443	0.40%
Available images with data in at least one queried tag	84,113	81.60%

Table 3 First five observations of our working data frame, shaped 68,085 rows × 5 columns

Camera manufacturer	Camera model	Exposure	Aperture	Focal length
Canon	Canon PowerShot S2 IS	1/640	4.0	72.0 mm
Panasonic	DMC-FX9	1/13	3.6	9.9 mm
Canon	Canon EOS 20D	1/250	11.0	560.0 mm
Nikon	NIKON D50	1/250	5.0	125.0 mm
Canon	Canon PowerShot SD600	1/320	2.8	5.8 mm
…	…	…	…	…

- Ultra wide (< 24 mm)
- Wide (24–35 mm)
- Normal (35–85 mm)
- Short telephoto (85–135 mm)
- Medium telephoto (135–300 mm)
- Super telephoto (+ 300 mm)

We also parsed over exposures to remove faux entries (e.g. a small number of older mobile phones reported infinite or zero values for exposure), and manually matched some camera manufacturers names (e.g. 'NIKON' and 'Nikon Corporation'). The consolidated data frame includes all values in all remaining tags for a total 68,085 entries, which is 66% of all images that comprise the Visual Genome (v1.2). An example of our working data frame is shown in Table 3.

Our analysis of EXIF data shows the clear dominance of DSLR over other types of equipment, with Canon and Nikon being the two major manufacturers combining for over 64% of all cameras, more than eight times the share of the third largest manufacturer, Sony, at 8% (Fig. 4).

From these, the ten most popular camera models all correspond to Canon EOS and Nikon DX systems, with the only exception of the Apple iPhone 4, at number nine. The most common camera in our dataset is Nikon's D90, an entry-level DSLR released in 2008, and the first model with video-recording capabilities. The second most popular is the semi-professional Canon 5D Mark II, released the same year, closely followed by the 7D also from Canon, released in 2009.

In terms of how these cameras were used, our analysis identifies large apertures *f* 2.8, 4, and 5.6 as the most popular, accounting together for 74% of photographs (Fig. 5).

For focal length, lenses between 35-85 mm are the most common, accounting for 50.7% of the images, with the least popular being the super telephoto, only used to take 1.6% of the photos in our dataset (Fig. 6).

Exposure was more evenly distributed between the extremes with the notable exception of 1/60, identified as significantly more popular than all other shutter speeds. This is possibly due to the common belief that this is the slowest shutter speed one can expose without needing a tripod.[14]

Footnote 13 (continued)

would need to extract the size of the sensor from Exif (assuming this is not given as the 35 mm equivalent), calculate the crop factor for each individual camera model, and then match it to the corresponding entry in the dataset. We did not have the time or resources to do this. However, through controlled manual sampling we identified entries that reported focal lengths consistent with two types of cameras widely available at the time these pictures were taken: 3G Mobile phones (~ 15 k entries, e.g. iPhone 4 to 5), compact and ultra compact point and shoot cameras (~ 12 k entries, e.g. Canon Powershot and Pentax Optio series). Based on this we compensated for these two groups by applying a weighted average crops factor of 7.6 and 4.8, respectively. From a similar sampling at the other end, it was apparent that this process not necessary for long focal lengths, which were mostly taken with full frame or APS-C or APS-H cameras, which magnify the image even more.

[14] This is commonly known as the "1/focal length rule". According to it, for a 50 mm lens (75 mm in most APS-C sensors), the longest exposure that still produces sharp images with hand-held cameras would be approximately 1/60. This is only a guideline, as many other factors impact sharpness: ISO, time of day, weather, and indeed how much one's hand shake. Still, as rule of thumb for DSLR aficionados, it might contribute to explain the popularity of this particular shutter speed.

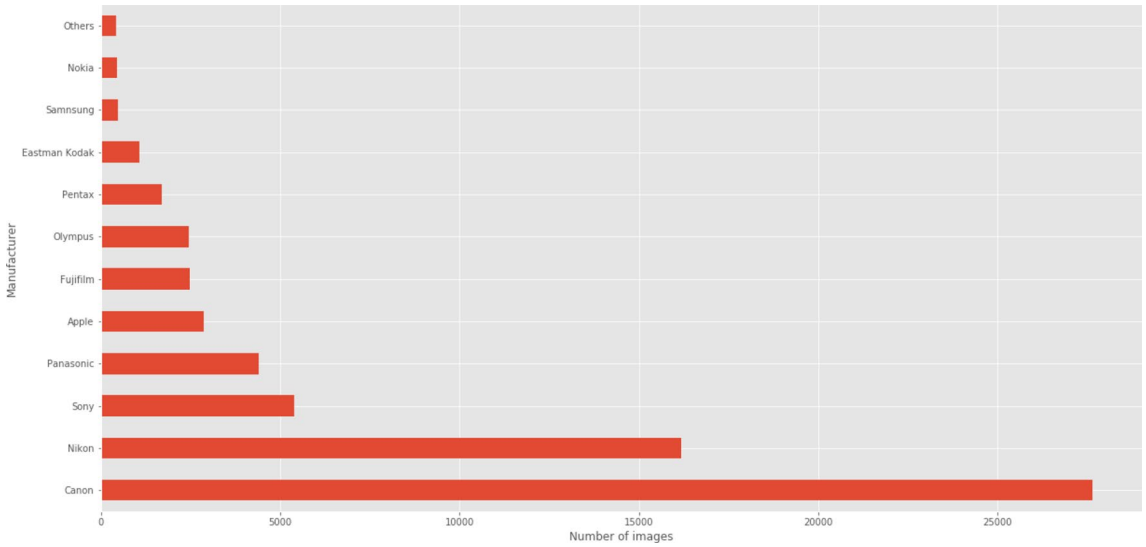

Fig. 4 Camera manufacturers of images in the Visual Genome

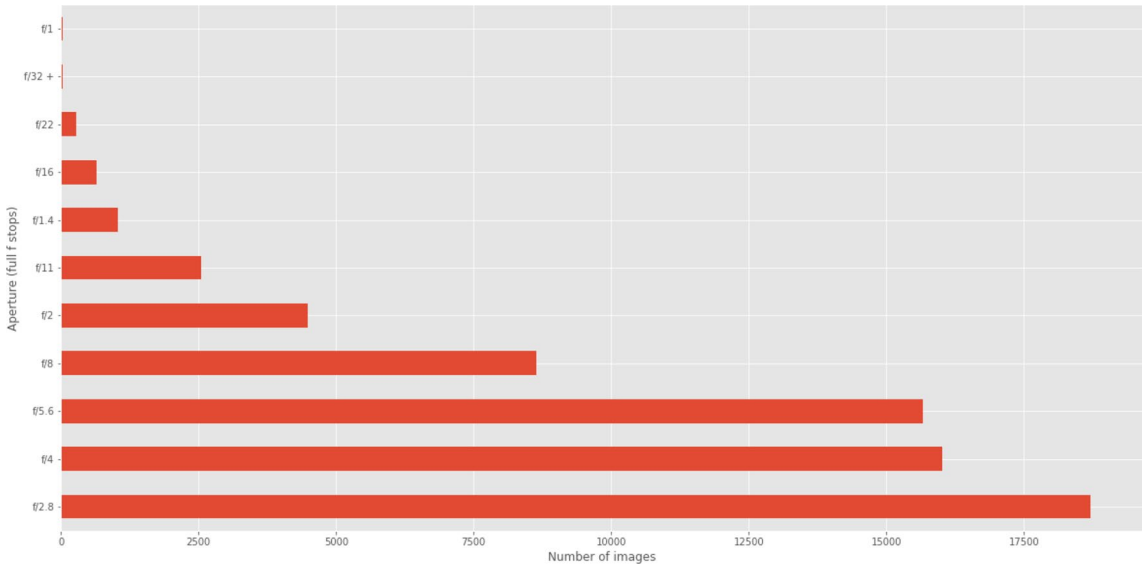

Fig. 5 Distribution of apertures in images from the Visual Genome

Figure 7 shows the ten most common combinations of aperture, focal length and exposure parameters in images in the Visual Genome, all of which are under direct control of their photographers.

These findings are consistent with the practices of a "proficient consumer" community of photo enthusiasts working with DSLR equipment.[15] These are generally non-professional photographers who nevertheless are willing to invest in a bulkier and more expensive camera and take the time to learn how to operate it manually. Users of the Nikon D90 are often recent converts migrating upwards from the

[15] Single-Lens Reflex cameras (both digital and analogue). This type of camera allows for interchangeable lenses and has a mirror system that allows the photographer to use a view finder to see through the camera lens in order to compose their photographs. When the shutter is pressed the mirror flips and the sensor or film stock gets directly exposed to light coming in through the lens. The acronym is often used to differentiate these cameras from point-and-shoot models, which are much smaller and have fixed (often retractile) lens, or from so-called mirrorless models, which do admit different lenses but do not have a mirror.

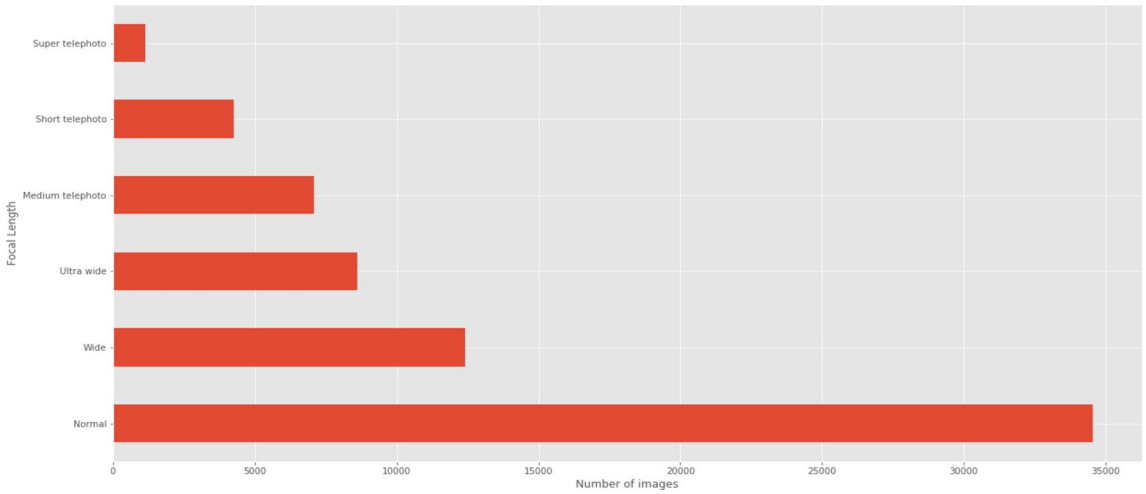

Fig. 6 Focal length categories in images from the Visual Genome

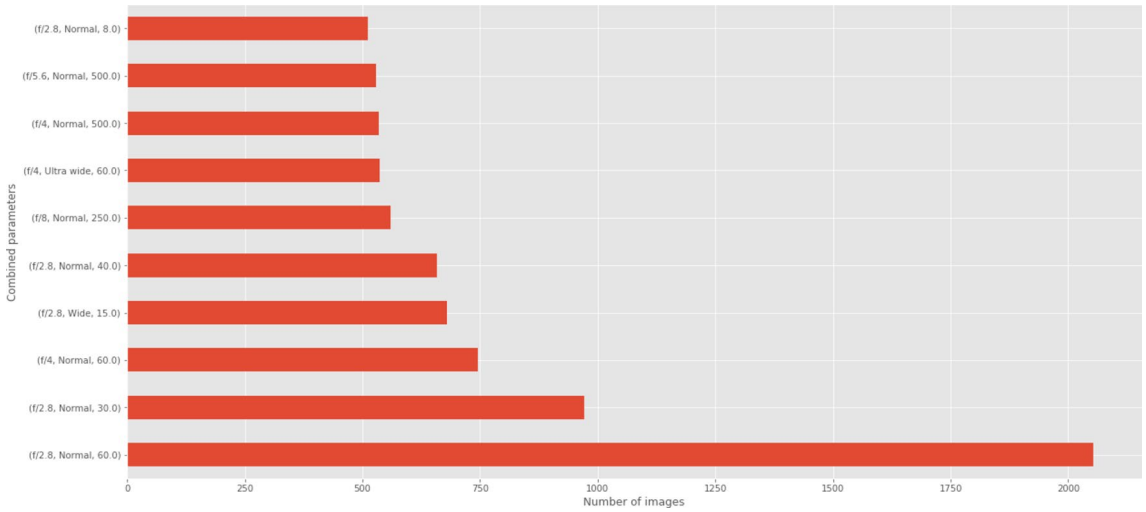

Fig. 7 Combined aperture, focal length and exposure* of images in the Visual Genome. *Exposure is given as the denominator of a fraction of a second, e.g. 250 is equivalent to 1/250, or 0.004 s

common point-and-shoot photography. Or they might also be more established and committed users of a Canon 7D, who probably own a few lenses already and might be close to going professional. This grouping is also supported by our smaller sample of lens data from the <Lens Info> tag, which shows inexpensive lenses that come bundled with cameras to be very popular, e.g. the 18-55 mm *f*/3.5–5.6 included in both Nikon and Canon starter kits (camera body + lens), but also include a few more expensive lenses (particularly on longer focal lengths, e.g. the 100–400 mm *f*/4.5–5.6L or the EF70-200 mm *f*/2.8L, both by Canon).[16] We believe these lenses overlap with professional practice and were probably acquired as second or third lenses for purpose-specific photography, specifically wildlife or sports, both of which featured heavily in a manual sampling we conducted over images taken with these two models.[17]

By identifying the dominant photographic practices of this community of DSRL enthusiasts in the Visual Genome, we show the *implicit optical perspective* mobilised in MbM. If one were to ask not about the accuracy in detecting

[16] We believe it is possible that Canon lenses on this range make more consistent use of the <Lens Info> tag.

[17] We looked at about 10% of the images taken with these two lenses.

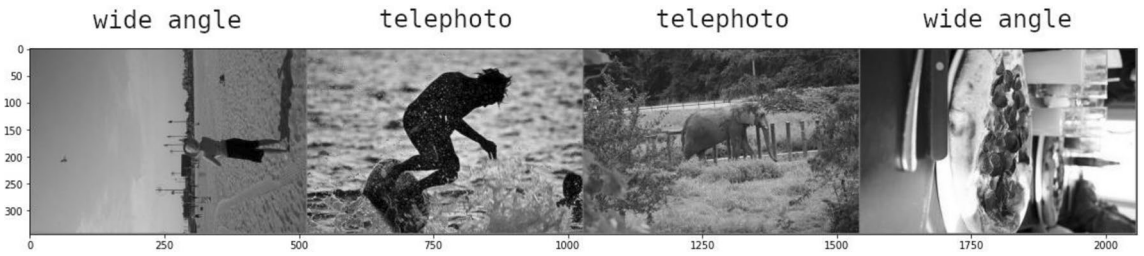

Fig. 8 Batch of four samples of inputs and labels

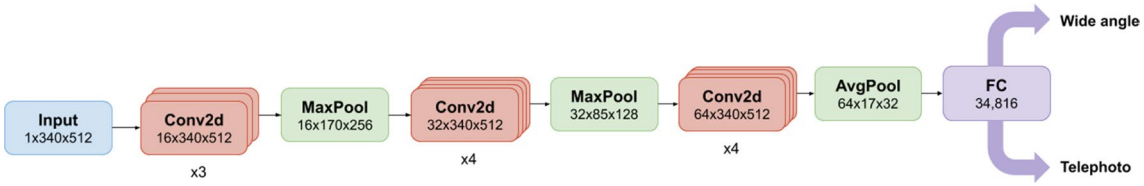

Fig. 9 15-layer Convolutional neural network architecture

what is depicted, but about the *latent camera* of this particular computer vision system, we could now reply with some degree of confidence that this perspective falls within the focal range of a 18–55 mm lens on a APS-C or APS-H camera; apertures between *f*3.5 and *f*5.6, and a likely exposure 1/60 s. Casting aside some of the other complexities of MbM for a moment, we could say that in general terms this was the lens through which the BBC archive was seen.

Today, DSLR photography of this kind is a somewhat dying practice, as sales of this type of camera have been steadily declining over the past decade (CIPA 2019). Everyday photographs are now taken with mobile phones and circulated through social media (Herrman 2018). However, while the equipment and the communities that supported this visual regime recede into history, lens aesthetics are anything but history. On the contrary, the standard of photography set by DSLR practitioners is now being reimagined under the logic of digital computation and mobile phones,[18] pursued through software and through AI (See for example: Yang et al. 2016; Ignatov et al. 2017).

With this in mind, we suggest turning computer vision to itself and asking whether it is possible to engineer a machine that tells us about the becoming of images; not only *what* they depict but *how*. If we concede that the "aboutness" with which we invest photographic images—including their epistemic advantage—is a function of the depicted no less than of the depiction modality, such an aesthetic machine, we argue, is as justified as one that distinguishes cats from dogs, or hot-dogs from other sandwiches. Could we not train machines to learn about optical perspectives as well as what these perspectives are used for at given times in history?

To close this article, we offer a prototype along these lines as a proof-of-concept, which is purposefully designed to be blind to what photographs are of; a type of vision that cares nothing about recognising objects, people or scenes, and is instead programmed to learn only about how its images were made and the visual perspectives they embody, in this case the focal distance of the lenses with which they were taken.

Using the EXIF dataset we assembled and the images from the Visual Genome, we trained a Neural Network to classify focal lengths and to distinguish between photographs taken with a wide angle lens from those taken with a telephoto. The class boundaries are drawn at under 24 mm for the former and over 135 mm for the latter. Each class was given little over 12,000 training samples (Fig. 8). The model was trained from scratch using a VGG-based convolutional neural network (Fig. 9).

Our results show test accuracy of 83% after fourteen epochs of training. We manually tested the model at this checkpoint by running inference on several photographs not contained in the Visual Genome to confirm it performed as expected in evaluating out of sample images. But we are only at the beginning of our work here. Without probing further into the model and conducting more systematic tests, it is difficult to know what exactly the neural network has

[18] Consider how large phone manufacturers like Huawei and Nokia partnered over the last decade with camera and lens manufacturers Leica and Zeiss, respectively, to produce multi-camera devices. In the case of the Huawei P series, such partnership was overtly marketed with the slogan "rewrite the rules of photography", in direct reference to its capacity to reproduce and control lens effects such as shallow depth of field and bokeh, which occur at longer focal distances and narrower apertures, and had until recently been the sole province of photographic cameras, most notably professional and semi-professional SLRs. See: https://consumer.huawei.com/uk/campaign/rewritetherules/

Fig. 10 Photo on the left was taken with a mobile phone (28 mm); photo on the right with a DSLR (340 mm)

learned from these images. One of our working hypotheses is that there are low-level features like the texture of bokeh or warmer green tones which might correlate strongly to longer focal lengths, since both the field view and speed of many of these lenses favours their outdoor use. In any case, our initial results already suggest that, with some exceptions such as irregularly shaped images from elongated panoramas, grainy images or images captured with optical zoom, the predictions of our classifier were reasonably accurate for photographs taken with either very long or very wide lenses. Figure 10 shows a comparison of two successfully classified images using this method. For the casual observer who sees these two images all at once, instead of counting them pixel by pixel, there are many apparent differences: one is the Shard in London, the other a baby orangutan in Borneo; one is a landscape, the other a portrait; one is a night scene, the other was taken in broad daylight. However, when it comes to the type of lens used to render these scenes visible, a *posteriori* knowledge might in fact be a task for which computer vision is much better suited. In particular deep convolutional networks can help with their progressive and content-agnostic abstraction of pixel relations.

Going back to MbM, we used our focal length classifier on frames from one of the mislabelled sections mentioned at the beginning (Fig. 11). Comparing the predictions outputted by the two systems, our 'telephoto' classification seems intuitively more accurate than MbM's 'reflection in a mirror'. This might be an extreme example but it points us to a fundamental problem that is sometimes overlooked in machine learning. Which prediction tells us more about the image? What kind of knowledge is implied by each, and when or why would we prefer one kind over the other?

5 Conclusions and future research

Initial results suggest that with more data, extended training and fine-tuning, a much more sophisticated lens classifier is possible. To our knowledge this has not been tried before, and therefore, we approached the design of the system with naive confidence, in the hope that others might be intrigued and improve upon it. Similarly, we believe there are many other possibilities beyond a binary classifier, even using this relatively small dataset, for example by looking at exposure as one of the immanent temporalities of computer vision systems—photographers not only manipulate the shape of light but also its speed. These need not be isolated dimensions nor indeed separate from existing approaches aimed at naming objects or people.

Fig. 11 Predicted class of (the whole) image on the left using our focal lens classifier prototype. Predicted label of (a region) of the image on the right in MbM

Our contribution is, rather, an initial and tentative answer to a much larger question: how can computer vision evolve from systems designed to name what is in the picture, to systems that approximate more precisely what we see in the picture? In this paper, we show from a computational aesthetics perspective how diversity in photographed subjects ought not to be confused with visual diversity, nor indeed bias with error. To be clear, our argument is not that Crawford's archaeology of datasets is not necessary, or that Harvey is mistaken when he states that a photo is not just a photo any more. It is rather that "just a photo" includes a whole field of meanings and technical mediations that are also encoded, abstracted and mobilised through machine learning and deep learning in particular. This is not to deny the digital dimension of these images, nor the latent computational powers of datasets, but to say that these powers are largely derived from the representational powers of photography. Therefore, our relationship with the photographic image underwrites our relation with computer vision more generally. From this perspective, a critical programme of computer vision, insofar as it is powered by this type of images, necessitates a techno-aesthetics of photography to explain how these images afford knowledge by distorting perspective, and how they can be seen as faithful representations beyond the factual events they depict. So here we must simply insist in incorporating insights from the study of the photographic and cinematic image in the technical milieu of AI to argue that meaning is not something that can be extracted from pictures alone, but that is instead co-constructed with their audiences through usage; photography is an imaging no less than an imagining technology, and so too must we learn to understand computer vision.

Acknowledgements Daniel Chávez Heras is funded by Mexico's National Science and Technology Research Council (CONACYT).

Compliance with ethical standards

Conflict of interest The authors hereby declare to not have any conflict of interest or competing interests. The dataset and code produced for this article is publicly available under MIT licence from: https://gitlab.com/chavezheras/shape_of_computervision. A training dashboard of the classifier model described in this article is publicly available at: https://cutt.ly/4taTWmt

References

Abell C (2010) The epistemic value of photographs. in philosophical perspectives on depiction, ed. Catharine Abell and Katerina Bantinaki. Oxford University Press, Oxford

Bazin A (1960) The Ontology of the Photographic Image. Translated by Hugh Gray Film Quarterly 13:4–9. https://doi.org/10.2307/1210183

BBC (2018) BBC—Public Purposes. https://www.bbc.co.uk/corporate2/insidethebbc/whoweare/publicpurposes Accessed 02 Oct 2018

Cavell S (1979) The world viewed: reflections on the ontology of film. Harvard University Press, Cambridge

Chávez Heras D, Blanke T, Cowlishaw T, Man D, and Herranz Donnan, A (2019) Seen by machine: computational spectatorship in the BBC television archive. In ADHO Proceedings. Utrecht, Netherlands

CIPA (2019) Camera and imaging products association: statistical data report. Sales and shipment report. Digital Cameras

Clark A, Chalmers D (1998) The extended mind. Analysis 58:7–19

Cohen J, Meskin A (2004) On the epistemic value of photographs. J Aesth Art Crit 62:197–210
Costello D (2017) On photography: a philosophical inquiry. Routledge
Cowlishaw T (2018) Using artificial intelligence to search the archive. BBC R&D Blog. https://www.bbc.co.uk/rd/blog/2018-10-artificial-intelligence-archive-television-bbc4 Accessed 28 Sept 2018
Crawford K, and Paglen T (2019) Excavating AI: The politics of images in machine learning training sets. AI Now Institute. https://www.excavating.ai/ Accessed 10 Sept 2019.
Cumberbatch G, Bailey A, Lyne V, Gauntlett S (2018) On-screen diversity monitoring BBC One and BBC Two. Media Monitoring. Cumberbatch Research group
Currie G (1999) Visible traces: documentary and the contents of photographs. J Aesth Art Crit 57:285–297. https://doi.org/10.2307/432195
JEITA Standards. Exchangeable image file format for digital still cameras: Exif
Harvey A, LaPlace J (2019) MegaPixels: origins, ethics, and privacy implications of publicly available face recognition image datasets. https://megapixels.cc/ Accessed 18 Apr 2019
Herrman J (2018) It's almost 2019. Do you know where your photos are? The New York Times
Ignatov A, Kobyshev N, Timofte R, Vanhoey K, Van Gool L (2017) DSLR-quality photos on mobile devices with deep convolutional networks. arXiv:1704.02470 [cs]
Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123:32–73. https://doi.org/10.1007/s11263-016-0981-7
Li F-F (2019) ImageNet 10th birthday party, september 21. The Photographers Gallery, London
Lopes DI (2016) Four arts of photography: an essay in philosophy. Wiley, Hoboken
Maynard P (1997) The engine of visualization: thinking through photography, 1st edn. Cornell University Press, Ithaca
Pasquinelli M (2019) Three thousand years of algorithmic rituals: the emergence of AI from the computation of space. e-flux 101
Peres MR (2007) The focal encyclopedia of photography: digital imaging, theory and applications, history, and science. Taylor & Francis
Phillips DM (2009) Photography and causation: Responding to Scruton's scepticism. The British Journal of Aesthetics 49. Oxford University Press: 327–340. https://doi.org/10.1093/aesthj/ayp036 Walden, Scott. 2005. Objectivity in Photography. The British Journal of Aesthetics 45. Oxford Academic: 258–272
Wieczorek M (2019) What I think about when I think about Focal Lengths. Medium. Accessed 28 Dec 2019
Yang Y, Lin H, Yu Z, Paris S, Yu J (2016) Virtual DSLR: High quality dynamic depth-of-field synthesis on mobile platforms. Electron Imaging 18:1–9. https://doi.org/10.2352/ISSN.2470-1173.2016.18.DPMI-031