PEER-REVIEWED ARTICLE

"WHY SO MANY WINDOWS?" – HOW THE IMAGENET IMAGE DATABASE INFLUENCES AUTOMATED IMAGE RECOGNITION OF HISTORICAL IMAGES

FRANCIS HUNGER

ABSTRACT | In the field of automated image recognition, computer vision or artificial 'intelligence,' the ImageNet data collection plays a central role as a training dataset. For the research project Training The Archive, which aims to make digital humanities methods available for the curating of art, the extent to which ImageNet influences the software prototype The Curator's Machine is discussed. The Curator's Machine is designed to facilitate the discovery of relationships and connections between artworks for curators. The text explains how ImageNet, anchored in contemporary image worlds, acts on contemporary and historical artworks by 1) examining the absence of the classification 'art' in ImageNet, 2) questioning ImageNet's lack of historicity, and 3) discussing the relationship between texture and outline in ImageNet-based automated image recognition. This research is important for the genealogical, art historical, and coding related usage of ImageNet in the fields of curating, art history, art studies and digital humanities.

KEYWORDS | Computer vision, convolutional neural networks, feature extraction, painting, artwork

1. Introduction

Artificial intelligence, machine learning, and computer vision are automation practices that hold the promise of generating new knowledge. What happens when framed art is detected as a television or a window? What does it mean when the drapery in Gothic paintings and sculptures is classified on the basis of the Ikea product catalogue? What happens when a painting by Lucas Cranach the Younger is processed on the basis of texture rather than outline in 'neural' or weighted networks?[1]

The present paper essentially revisits Dominik Bönisch's article *The Curator's Machine: Clustering Museum Collection Data by Annotating Hidden Patterns of Relationships Between Artworks* (2021). Bönisch focuses on a concrete software prototype, 'The Curator's Machine', which was developed in the frame of Training the Archive and is documented as open source.[2] Situated in the field of computer vision[3] Training the Archive investigates "a machine-supported, explorative (re)discovery of links within museum collections" (ibid., 1). Building on the former, this article will raise questions relevant to the interplay of the fields of art, curating, art history, digital humanities, and computer vision. Basic knowledge of how weighted networks and automated image recognition function is required for understanding this text.[4] In the first section, the image data to be processed are identified as 'operative' images. While representative images are aimed at image content that is explicitly made by people for people, operative images exist to be processed by machines. Next, the process of operationalising the images in those large image collections that train the artificial weighted networks, such as VGG16 or InceptionV3, is investigated. A further section is devoted to the absence of art in the image data collection ImageNet, which is one of the training standards for today's computer vision. The following section discusses the contemporaneity of operative images in ImageNet in relation to the historicity of painting, graphics, and sculpture in museum collections. This is followed by a discussion of texture and outline. How do texture and outline correspond

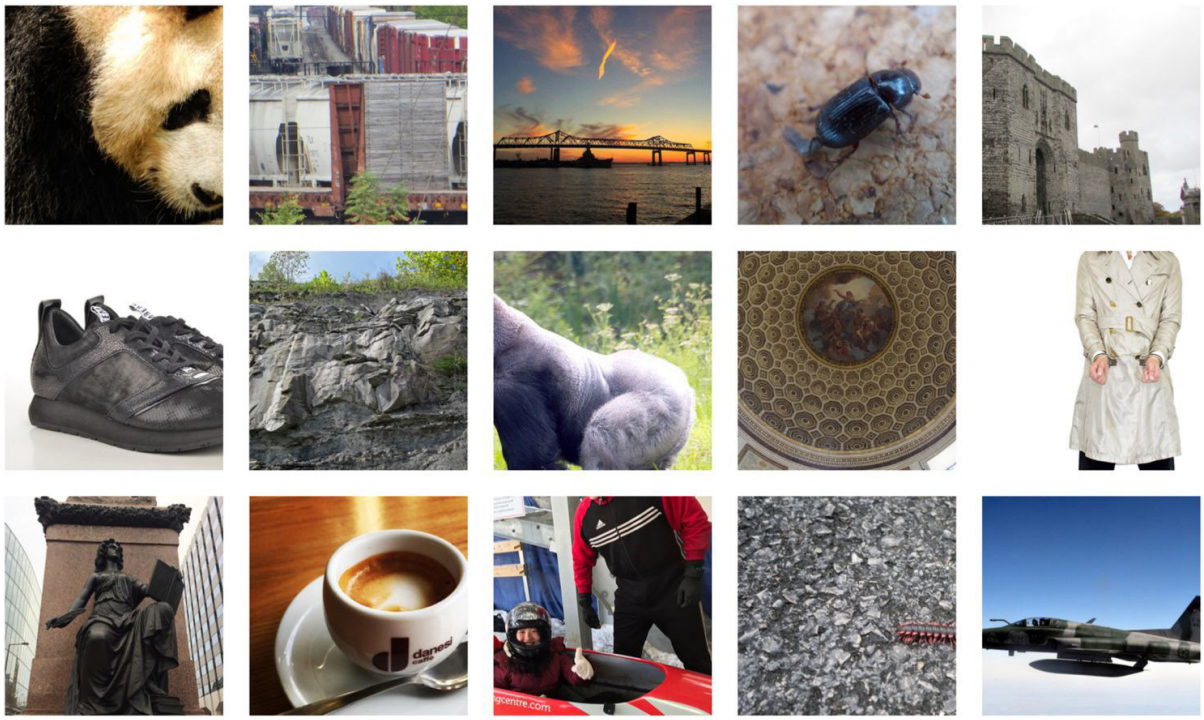

Figure 1: Operative Images 1 – training data from ImageNet (author; copyright for each image belongs to the respective author, 2021).

with the contemporary training images in ImageNet and the historical image material?

The focus of this working paper will not be on classification problems, such as 'bias', which have been discussed elsewhere (Noble 2018; Crawford and Paglen 2019). Rather, image-'immanent' problems are to be pursued here, for example, the question of how pre-processing, image formats, and historical use of form affect the later classification performance of trained weighted networks.

Why is the focus on the ImageNet image database? ImageNet is currently very widespread and is used in numerous research approaches as a benchmark for the effectiveness of weighted networks. While other image collections of similar or greater scope are currently emerging, such as Google's open-source Open Images Dataset, ImageNet is also evolving and drawing on the advantage of having been a pioneer in the research field of computer vision.

ImageNet, as a 'Canonical Training Set' (Crawford and Paglen 2019), comprises 14 million images annotated with labels to describe their content (Li et al. 2009; Hinton, Krizhevsky, and Sutskever 2012).[5] The images range from amateur and professional photography, most of which was downloaded from the Flickr photography platform, to product and stock photography taken from commercial websites. Offert and Bell, for instance, advocate a critical analysis of ImageNet:

> "A more broad critical approach would be the analysis of highly common datasets like ImageNet, which are not only used 'as is' in real-life classification scenarios but even more often used to pre-train classifiers which are then fine-tuned on a separate dataset, potentially introducing ImageNet biases into a completely separate classification problem." (Offert and Bell 2020, 9).

Weighted networks usually aim to classify objects within an image. However, The Curator's Machine is not intended to detect image content, but rather similarities between images and their features. In the course of the Training the Archive project, the final classification component of the weighted network is therefore switched off (see Fig. 3). Instead, the features calculated up to that point for each image from the input data are saved for further processing using the weights pre-trained with ImageNet.[6] The question then naturally arises as to what influence the ImageNet image database has on feature extraction.[7] Before this question can be pursued, however, the status of those images we are referring to must first be clarified. Are they images at all? Are they data? The next section explores the extent to which the images channelled through computer vision systems are 'other' images.

2. Between Representative and Operative Images

The images from image archives, which are being processed through weighted networks in the course of the Training the Archive project, change their status from 'representative' to 'operative' images. In the following, representative refers to non-operative images, i.e., images that are designed to be interpretable, that convey ideas, regardless of whether they are representational or abstract images. Images are operative when they enable a series of automated operations, for example identification, control, visualisation, and recognition (Broeckmann 2016, 128–134). Operative images are embedded as encoded data in process chains. Their primary purpose is the automation of knowledge-building procedures. Since their purpose is not representation but operation, operative images are subjected to pre-processing after capture or scan – a series of image manipulations for the purpose of preparing the dataset for further processing. In the following paragraphs, we will reflect on the homogenisation of operative images.

The programmer and artist Nicolas Malevé worked out how the photographic image is used as a homogenising procedure. In the preparation phase, the incoming image data should be as formatted as possible and comparable to each other:

> 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 (Malevé 2020, 6).

Stated more generally, in order to achieve homogenisation, operative images are subjected to labour-intensive pre-processing, in the course of which exposure ratios, colour space, distortions, size ratios, horizontal alignment, and similar parameters are standardised (Brownlee 2019). In addition to the pre-processing, the images are provided with partially automated and partially manually created annotations, so-called metadata, such as time and place of capture, equipment used, names of the editors, and editing status.

The existing digital image archives have been recorded with photo-optical sensors of cameras, scanners, and similar imaging devices, e.g., infrared, UV, MRI or radar (cf. Amat and Casals 1992, 3–8; Parikka 2021, 185–188). Every archive works with very different camera techniques and lenses, which limits the inter-operability of images from one archive to another. Even within the same archive, the comparability of operative images is often questionable. Their comparability depends on the skills of varying photographers, the various devices that are replaced over time by new equipment and software due to wear and tear, changing informational needs that are subject to institutional fluctuations, and changing standards in the field of knowledge (e.g., art history or digital humanities). Since The Curator's Machine does not yet work with its own datasets – these are in preparation – but instead accesses operative images provided by third parties, the project inherits the homogenisations inscribed there.

In summary, a sequence of image manipulations makes the images increasingly operative. This occurs in an independent step during the basic digital capture of a collection, which thereby becomes a collection of image data:

1. Set up the object to be photographed in a neutral picture environment (light, background).
2. Photograph or scan.
3. Annotate and enhance with metadata.
4. Complete general image post-processing with the aim of a homogeneous image data collection. This may also include the cleaning up of 'ageing phenomena.'
5. Prepare for long-term archiving and, if applicable, publication of the collection.

If a research project, such as Training the Archive, accesses such a collection of images in order to process them further in the course of computer vision, additional steps take place:

6. Clean up the image data set (e.g., removing duplicates, removing colour wedges, cropping edges).
7. Make needed format adjustments of the image data, including cropping to square image format for processing by weighted networks, such as InceptionV3, or VGG 19.

Already before entering the weighted networks of Computer Vision algorithms, a variety of formatting techniques have affected the state of the 'original' images.

The concept of image reaches its limits here, since we are dealing with datasets that have already been subjected to calculations during pre-processing before they are even fed into the actual computing apparatus, the weighted network:

> […] we have to look beyond the image – or even the sensor – as a stand-alone unit, and instead understand that the image is, at best, an interface (Bratton 2015: 220–6; Andersen and Pold 2018) that allows a kind of access to other scales of infrastructural action that mobilise multiple kinds of knowledge of large-scale, dynamic systems […]. (Parikka 2021, 203).

With regard to weighted networks, it is important to distinguish between two different forms of operative images:

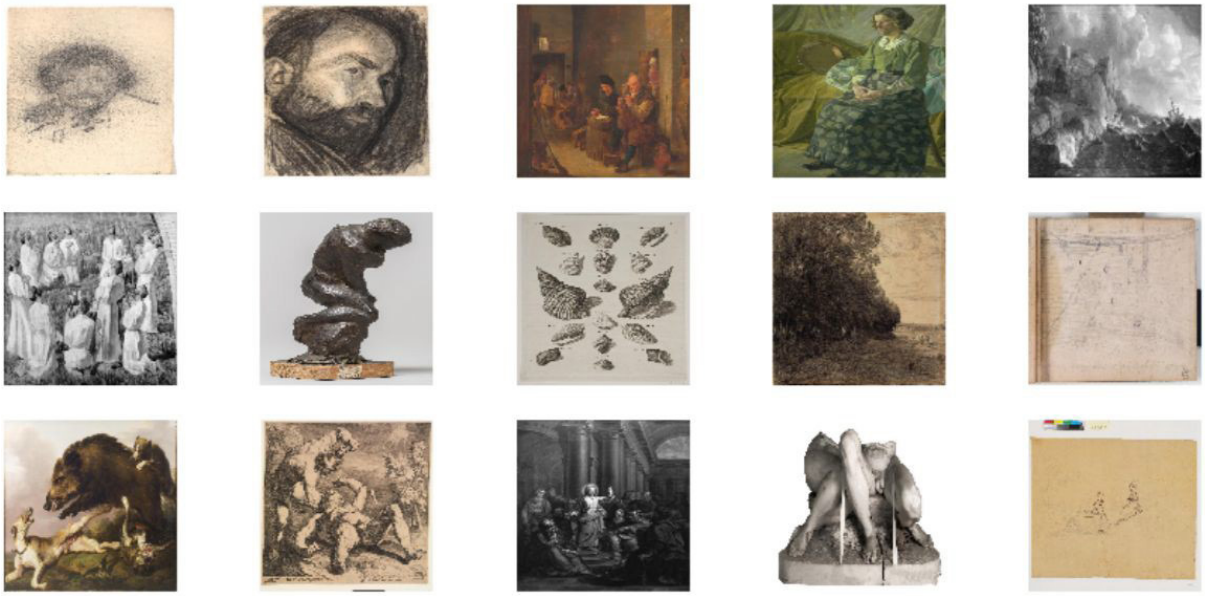

Fig. 2: Operative Images 2 – Input data from the National Gallery of Denmark Statens Museum for Kunst, square image sections (Bönisch 2021).

1. Operative images that train the weighted networks (Fig. 1), i.e., that inscribe weights on the nodes in the first place. The weighted networks InceptionV3, BiT/m-r152×4 and VGG19, used for the first prototype of The Curator's Machine, are based on training using the operative images of ImageNet.[8] These operative images are subsequently referred to as 'training data' in the text.
2. Operative images of an art collection to be classified by weighted networks (Fig. 2). The first prototype of Training the Archive was based on input data from the National Gallery of Denmark Statens Museum for Kunst. They were selected due to their free availability and the "high-quality data sets, e.g. with regard to the existing image resolution, data variance, and amount of meta-information" (Bönisch 2021, 22). In contrast to the first point, the input data are images of the collection items that have been post-processed and, as far as possible, standardised and idealised. These operative images are subsequently referred to in the text as 'input data.'

For operationalising in computer vision libraries such as Pytorch or Tensorflow/Keras, pixel counts of only 512×512 pixels (MobileNet V3 Large-M), 299×299 pixels (InceptionV3), 224×224 pixels (Resnet50, VGG16, VGG19) and others in square format have prevailed for reasons of processing efficiency.[9]

The current inaccessibility of digitised data and the great complexity of the project pragmatically justify the current approach of working with collection data from the Statens Museum for Kunst. However, the image data collection of the Statens Museum for Kunst, which is focused on painting, graphic art and a few sculptures and installations, reinforces the problematic normative of art as a two-dimensional medium. It is therefore also the task of Training the Archive to find strategies for ephemeral, multimedial, and conceptual contemporary art.

The operative images used here currently refer primarily to two-dimensional works as input data. With this limitation, it is now necessary to ask what computer vision can achieve as a curatorial tool for a limited body of works (paintings and drawings).

3. The Absence of 'Art' in ImageNet Pre-Trained Weighted Networks

Recent studies have shown that pre-trained weighted networks[10] sometimes do not classify art as such. With the help of the Wolfram Alpha platform, the artist Rosemary Lee examined the first 100 hits for the keyword 'abstract' in the image database of the Metropolitan Museum of Art. Lee found that 98% were not classified as art (Lee 2020, 92).[11] Pereira and Moreschi come to similar conclusions, and not only in relation to a corpus of abstract paintings. Computer vision interprets art primarily as everyday objects: "These readings invite us to see works of art in a way that is disconnected from the idea of authorship" (Pereira and Moreschi 2020, 6).

Based on these observations, the question arises whether and how art occurs in the underlying training data of the pre-trained weighted networks. For this purpose, the now

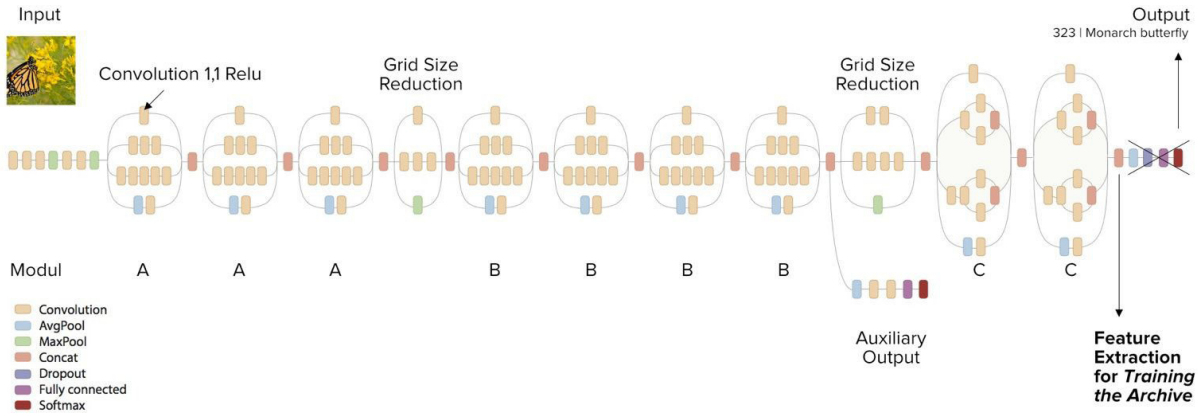

Fig. 3: Scheme of the Google InceptionV3 architecture. Each rounded rectangle represents a convolution or other mathematical function. The concatenations separating the modules are used to reduce the features. See color legend: concat (scheme based on Google https://cloud.google.com/tpu/docs/tutorials/inception, accessed June 28, 2022).

widely used data collection of operative training images ImageNet Fall 2011 (Li et al. 2009; Hinton, Krizhevsky, and Sutskever 2012) was examined. The annotations of the 15 million operative images with 1000 classifications collected in ImageNet originate from a psycholinguistic system, which was created for the Wordnet database at Princeton University from 1985 onwards. In Wordnet, verbs and nouns (such as chair, child, art) are given IDs and linked to each other, resulting in chains of affiliations, so-called synsets (synonym sets). The Wordnet synsets are used for the ImageNet image data collection to annotate images and attribute semantics. The concepts 'art' and 'painting' and 'picture frame' are present in Wordnet and should in principle also be addressable in ImageNet.[12]

Does the training data of ImageNet 2011 contain operative images that relate these categories? My investigation shows that none of the 15 million images in the ImageNet 2011 training data appear to be labelled as 'art' or 'painting' or 'frame.'[13] This does not exclude the presence of art in ImageNet 2011. However, art is not labelled as such. Does this pose a problem for the first prototype of The Curator's Machine? The prototype does not currently use label recognition but compares the image material on the basis of similarities. This is carried out using so-called features (mathematical vectors), which detect certain patterns within the images.

Figure 3 shows up to which point the ImageNet-trained network Inception is used in the process. In addition to InceptionV3, The Curator's Machine similarly uses the ImageNet-trained network BiT/m-r152×4 and individual modules from VGG19 in two variations.[14] InceptionV3 and BiT/m-r152×4 are pre-trained in The Curator's Machine with the image database ImageNet.

Another guiding research question has been: what do these features look like, and can they be depicted? In module A1, for example, there is a convolution in InceptionV3 (1.1 Relu, see Fig. 3), which responds to curtains or drapery. Since drapery played a major role in Gothic figurative representation, it will briefly be pursued here. With the help of the tool OpenAI Microscope,[15] parts of weighted networks can be visualised. There, we searched for drapery images and identified a specific convolution – Unit 45 – that activates corresponding patterns (Fig. 4, Fig. 5, Fig. 6).[16] These patterns correspond to the mathematical features that were trained using the ImageNet database.

For The Curator's Machine, as described above, the features from three weighted networks are merged before the images from the SMK collection (Fig. 2) are applied to them as input data. For each weighted network pre-trained on ImageNet, the images are entered as input data. Then, for each image, the network-specific features are extracted, and these are then merged at the end to build the latent space from them, whereby the features were once again reduced by mathematical means before building the space.

Computer vision practitioners currently assume that ImageNet-trained features generalise sufficiently when specifically trained features are added in later layers (Yosinski et al. 2014; Huh, Agrawal, and Efros 2016, 6; Kornblith, Shlens, and Le 2019). Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros found that the absence of classes showed little impact on the generalisability of features in ImageNet (ibid., 7). In contrast, the research team of Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah shows from their examination[17] that the 'neurons' or nodes of the weighted networks replicate the ontology of ImageNet and WordNet, respectively: "[…] it seems as though the neurons appear to arrange themselves into a taxonomy of classes that appear to mimic, very approximately, the ImageNet hierarchy" (Goh et al. 2021). Their findings demonstrate that further research is needed, because there is a lack of certainty that ImageNet pre-trained weighted networks remain unaffected by the absence of 'art.'

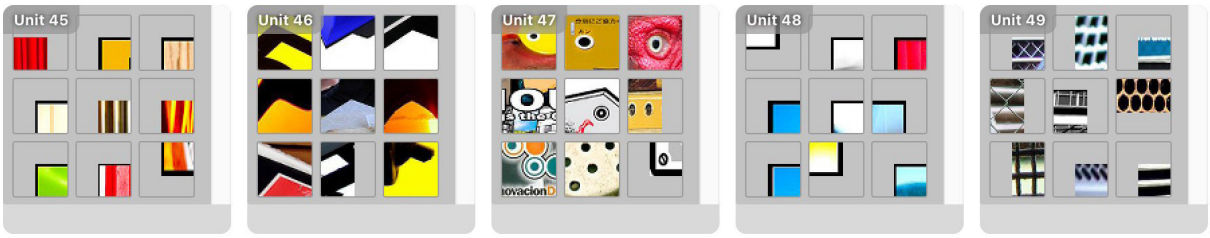

Fig. 4: Unit 45, (first tile) of Convolution 1.1 Relu reacts in an ImageNettrained InceptionV3 to curtains, among other things. The illustration shows examples of images from the ImageNet database (screenshot, OpenAI Microscope, 2021).

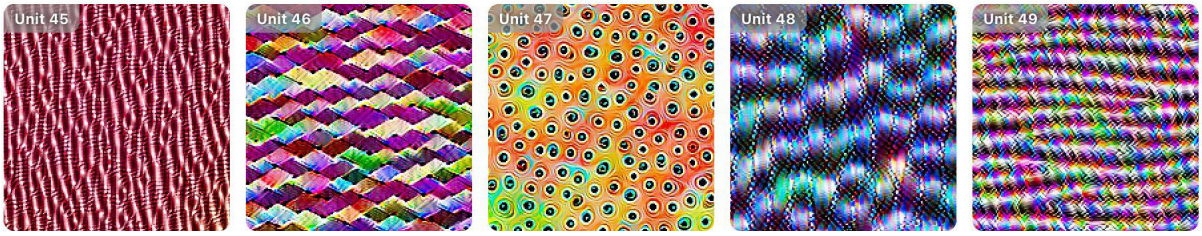

Fig. 5: Unit 45 (first tile) - This representation using the Deep Dream algorithm shows the patterns of specific units in the convolutions of the weighted networks. The first tile corresponds to the "curtain" pattern (screenshot, OpenAI Microscope, 2021).

In their text "Perceptual Bias and Technical Metapictures: Critical Machine Vision as a Humanities Challenge", Offert and Bell show that data bias can also occur in pre-trained networks, which is exactly the case of The Curator's Machine. Their example is the occurrence of fences. These should represent a chain-link fence structure. They analysed 1300 images from the 'fence' category in ImageNet and found that 1% to 5% of these training images show people confined behind fences and conclude: "images of people behind fences will appear more fence-like to the classifier" (Offert and Bell 2020, 9). Applied to our example, it can only be assumed at this point what kind of bias occurs with drapery. This requires further investigation.

In relation to the prototype of The Curator's Machine, we can therefore assume that the absence of art does not have any influence on the generalisability of the features. However, the problem of the absence of art in ImageNet pretrained weighted networks is not satisfactorily settled yet. This question could only be solved by tests focused on the domain of art. First, a training data set for art would need to be developed. This would include extended discussions between art historians, artists, and data scientists about which objects to include in such a training set. Subsequently, the procedure of 'backpropagation' during feature transfer allows the lower layers, for instance the first two A-modules (according to Fig. 3), to be fine-tuned based on the upper, transferred layers, so that the lower layers would generalise better with regard to the domain 'art.' This procedure could be considered in addition to using the lower, pre-trained layers (Yosinski et al. 2014, 6).

Art is not annotated as a category in ImageNet 2011 and if it is, it is only marginally present. Computer vision projects in the digital humanities that use pre-trained weighted networks for classification should include the genealogies of ImageNet 2011 in their considerations. For Training the Archive, it is necessary to ascertain whether back-propagation of domain-specific layers can change the lower layers of pre-trained networks. Research to date on the effects of pre-trained networks with features that continue to be used is incomplete and, therefore, comes to contradictory conclusions. For Training the Archive, this means that appropriate analyses must be carried out with reference to one's own datasets.

4. Lack of Historicity

What is striking about the use of ImageNet and other publicly available training data, such as Open Images Dataset and Microsoft COCO/Azure, is their lack of 'historical memory.' The input data for the first prototype of The Curator's Machine—the paintings and drawings of the Statens Museum for Kunst—are mainly from Europe, dating from the 15th century to the 20th century.

Over this long period, shifts in the content of the pictorial subjects and the modes of representation are significant. For example, the depiction of the body in Gothic painting differs from today's predominant body images by emphasising the length and extension of the limbs as an expression of courtly elegance. The complexity of the drapery had its very own meaning in the Gothic period, which is rarely found in today's image data.

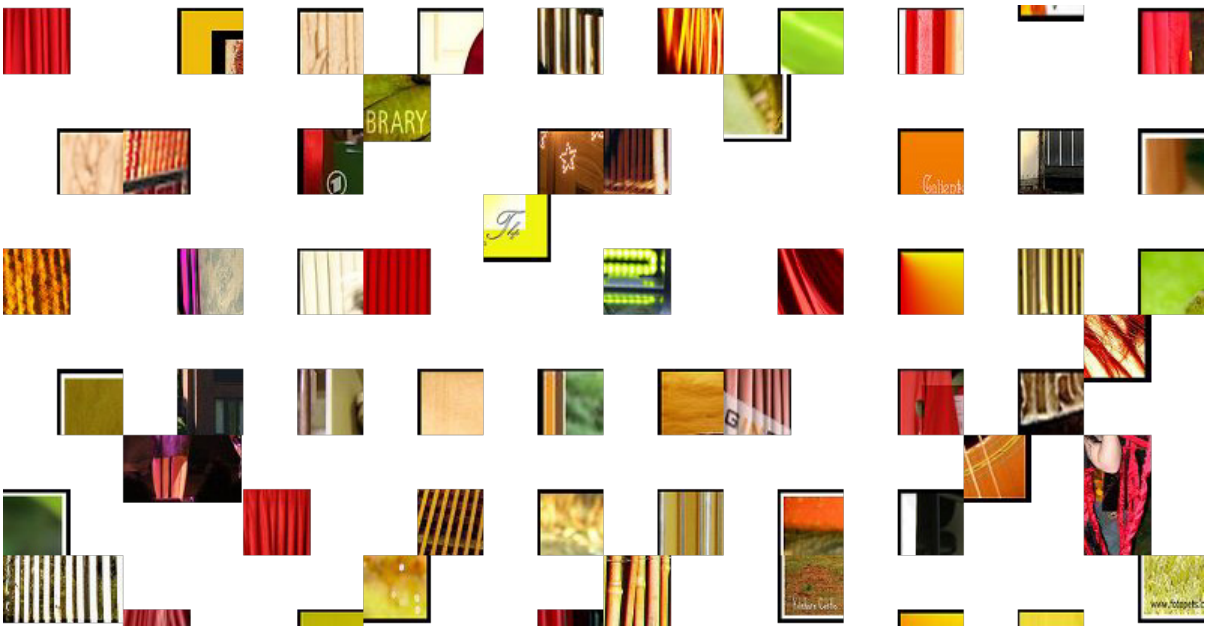

Fig. 6: Zoom on Unit 45: Closer inspection shows that curtains (red), but also other striped structures are activated, such as serially arranged fence slats or bamboo canes. The illustration shows examples of images from the ImageNet database (detail, screenshot, OpenAI Microscope, 2021).

These points illustrate specific historicities of the input data. These now encounter the contemporaneity of the ImageNet training data used for pre-trained weighted networks. The training data was downloaded from the Internet at various points in time from 2010 onwards, by querying search engines for images in five languages using the Wordnet synsets (Li et al. 2009). A large proportion of the images came from the photography platform Flickr. The labels of the ImageNet images were annotated by precarious click workers using the Amazon Mechanical Turk platform.

Objects photographed and classified from the 2010s onwards thus now train a weighted network that is supposed to process input data relating to the historical use of forms since the 15th century. The underlying problem is also recognised by the machine learning community itself: "These results suggest that classifiers based on modern machine learning techniques, [...] are not learning the true underlying concepts that determine the correct output label. Instead, these algorithms have built a Potemkin village" (Goodfellow, Shlens, and Szegedy 2015, 2).[18]

This problem becomes apparent in three current artistic projects. *What the Machine Saw* (2019) by John Stack labels a series of images from the collection of the Science Museum Group with the help of the Amazon Rekognition service (Fig. 7). These labels, automatically assigned by Rekognition and based on trained weighted networks, are juxtaposed with descriptions from the metadata to the images. The metadata was created by museum staff when the objects were added to the collection (Stack 2019). The work shows the difference between automated classification based on the statistical models of machine learning and the annotations created by humans as metadata.

A second artistic project comes from Philipp Schmitt. *Declassifier* (2019-2020) overlays specially photographed everyday scenes from New York with the images used for training the underlying Microsoft COCO dataset (Fig. 8). The example shows a Manhattan street scene with passers-by. If you move the mouse over an object frame (violet), which marks an object recognised by Computer Vision, one of the original photos from the training dataset appears. In addition, the authorship of the training images, which is ignored in the COCO dataset, is again indicated by stating the author, title, and file name in a white information box (Schmitt 2019). Schmitt's project impressively demonstrates the connection between training data and input data and shows how various spatial, temporal, and topological orderings collide with each other.

In the third project, *Recoding Art*, the artists Gabriel Pereira and Bruno Moreschi examined a portion of the collection of the Van Abbemuseum Eindhoven with 654 images (fig. 9). All images from the collection are cropped, lit, and colour optimised, making them well suited as operative images. However, size information and other metadata are lacking. With the help of a self-programmed software tool, part of the larger project *Recoding Art*, Pereira and Moreschi investigated how artworks are interpreted as everyday objects through

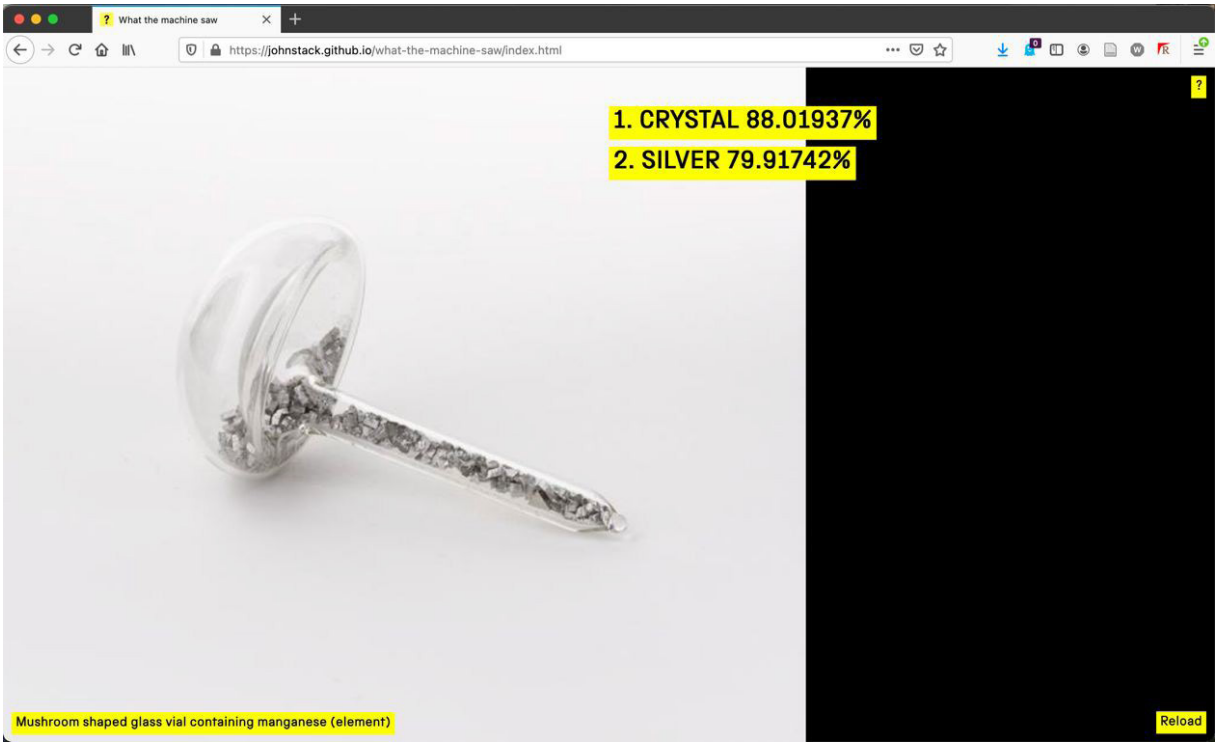

Fig. 7: John Stack, What the Machine Saw (2019): A mushroom-shaped glass vial filled with manganese is recognised as crystal or silver (screenshot).

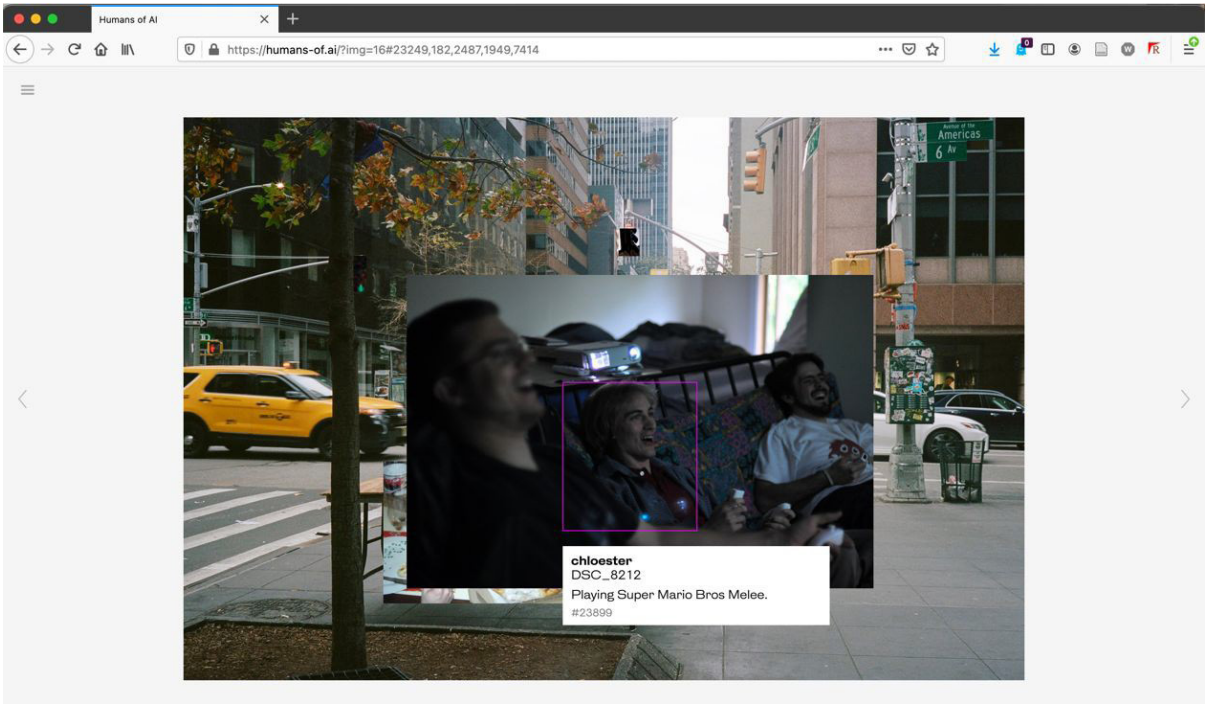

Fig. 8: Philipp Schmitt, Declassifier (2019-2020): A photo of a street scene in Manhattan with people is overlaid with images used to train the object recognition 'person,' in this case a picture of 'chloester' with the title "Playing Super Mario Bros Melee" (screenshot).

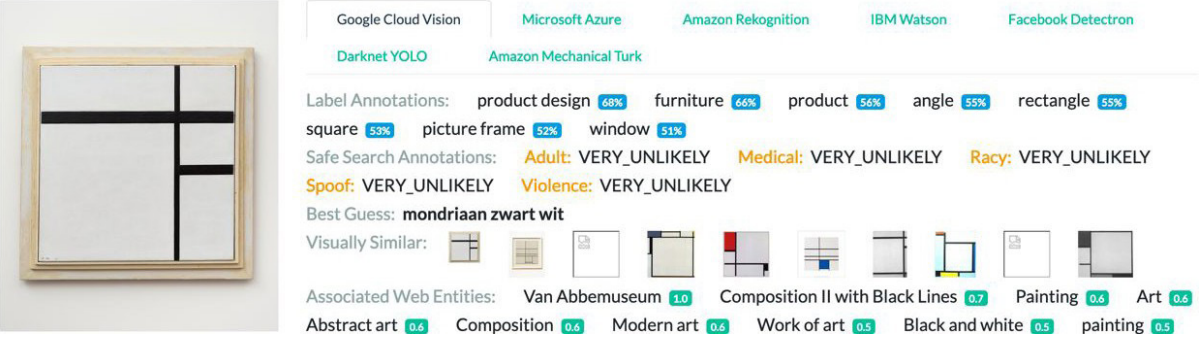

Fig. 9: Gabriel Pereira and Bruno Moreschi, Recoding Art (2021): Google Cloud Vision detects Piet Mondrian's "Composition en blanc et noir II", 1930 (screen-shot).

computer vision. *Recoding Art* outputs the interpretation for each individual image through a series of APIs: Google Cloud Vision, Microsoft Azure, Amazon Rekognition, IBM Watson, Facebook Detektron, and Darknet YOLO (Pereira and Moreschi 2020, 2).

Google Cloud, Amazon Rekognition, and IBM Watson are comparatively good at detecting 'art' or 'painting,' presumably because they also involve metadata and search results. The other services, detect individual objects (e.g., 'person,' 'table,' 'tree') with varying degrees of success, but do not recognise the concept of 'art' as such, or only with low probability. Google Cloud Vision is able to detect sculptures as such (e.g. Christos and Jeanne-Claude's *Wrapped Armchair* and Ernst Barlach's *Teaching Christ*) and non-figurative abstract paintings (Fernand Léger *L'accordéon*) as 'paintings.' From this observation, the question arises as to what distinguishes Google Cloud Vision from the other APIs. It can be assumed that the recognition applies to works related to the Google Arts and Culture Project, which has digitised numerous artworks and categorised them into 13449 artists, 240 media, and 117 art movements.[19]

With regard to the historicity mentioned above, Moreschi and Pereira state that:

"In at least one of their results, the vast majority of the works (almost 90%) were read as consumer products easily found in department stores" (ibid., 6). This finding should not be underestimated in terms of the relationship between the contemporary training data and the historical input data of the image collections to be examined. The authors state that a 'capitalist logic' (ibid., 12) is at work behind the currently trained networks that reproduces a corresponding normativity.[20]

The three examples given here, demonstrate the problematic ahistoricity of computer vision. If a 'drapery' (ImageNet-ID: 03237826) in a 15th century painting can be classified, then it is because drapery was trained with an operative image from the (metaphorical) 'Ikea product catalogue.' This suggests that a whole series of shapes, textures, and objects are not part of the training because they do not occur in today's product world, or are ahistorical because their former meaning does not correspond to today's meaning.

In the Gothic period, pronounced drapery folds were stylistically characteristic of sculptures and paintings (Sauerländer 1970). Art history distinguishes between "Folds, cascades of folds, trough folds, omega folds, parallel folds, pipe folds, bowl-like folds, conical folds, V-shaped folds, Y-shaped folds and zigzag-style folds" (Kunsthistorisches Institut der CAU Kiel 2019, my translation). Such taxonomies have little to do with those folds of curtains and drapes that exist today as commodity-based products.

However, if it is a question of 'decolonising' the curatorial perspective, which does not proceed solely in a strictly art-historical manner, new associations within the images can arise. Moreschi and Pereira suggest that detections through computer vision should not be perceived purely as a problem. Instead, ignoring authorship and historicity would open up potential for eliciting new narratives of art history through novel chains of association that reach across different geographical regions and temporal periods (Pereira and Moreschi 2020, 17). That the neural networks contrast art-historically 'valorised works' (Groys 2004) with similar amateur artworks disregarding the canon, would open the way for the perception of art beyond the museum (Pereira and Moreschi 2020, 20).

The ahistorically and ageographically[21] trained neuronal networks enable interwoven, post-humanistic human-machine figurations that create space for new epistemic processes. The aim of Training The Archive is to allow this decontextualisation in a first step in order to subject it to a new human evaluation in a second step, as a collaboration between curator and machine. The ahistoricity of the training data in relation to the input data was pointed out.

Fig. 10: When the texture of the skin of an Indian elephant is transferred (a) to a cat, (b) the texture-shape conflict leads to (c) the cat being recognised as an elephant (Geirhos et al. 2019, Fig. 1).

5. Texture and Outline

One aim of the first prototype of Training the Archive is to provide exploratory visual representations such as cluster analysis, gridplot, or scatter plot for the input data. These can additionally be trained by human input, as described in the section *Triplet-Loss-Function* by Bönisch (2021).

Texture and facture represent important art-historical and curatorial evaluation criteria. However, they not only determine the representative, but also the operative dimension of images. How do they affect the aspired representations in cluster, grid, and scatter plots? Texture and facture enter an interplay with specific medialities of historical images. Not only do the medialities and qualities of the picture carriers used, such as paper or canvas, pigments, colors, and inks, change over the centuries, but ageing processes, such as yellowing, darkening, fading, staining, and craquelure, also occur. Texture and facture have consequences for reception depending on the ageing and treatment process.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel have shown by experiment that ImageNet-conditioned weighted networks prioritise 'texture' over 'shape': "ImageNet object recognition *could*, in principle, be achieved through texture recognition alone" (Geirhos et al. 2019, 2). They term this "principle texture bias." (ibid.) For example, they filled the outline of a cat with the texture of an elephant. This image was reliably recognised as an elephant by weighted networks (Fig. 10).

From these results, they conclude that texture bias is one of the shortcuts not uncommon to computer vision and pattern recognition. The algorithm-data-systems are optimised to deliver a human-validated result by taking the shortest, i.e., most mathematically-economically optimised solution path: "If textures are sufficient, why should a CNN [Convoluted Neural Network] learn much else?" (ibid., 9).

How could texture bias be dealt with? Shortcuts are so common in weighted networks that even apart from the particular examples of age spots and craquelure, they need to be expected (Geirhos et al. 2020, 2f.). Identifying shortcuts procedures to mitigate these can be employed, building on existing research. Geirhos et al. show that a modified ImageNet trained weighted network with stronger emphasis on shape, mitigates texture bias: "We show that the texture bias in standard CNNs can be overcome and changed towards a shape bias if trained on a suitable data set" (Geirhos et al. 2019, 3).

As a procedure for correcting texture bias, they suggest using style transfer to generate a specific ImageNet image data collection as training data.[22] For the style transfer, a series of styles are passed to the training images using the AdaIN method (Huang and Belongie 2017), resulting in an emphasis on shapes. By using the same manipulated image from the ImageNet database in various styles for training, the authors were able to ensure that shape, rather than texture, was inscribed in the weights of the trained network (Geirhos et al. 2019, 5).[23] Why all this effort? It is known from human neurophysiology that people recognise images primarily on the basis of shape.

The first prototype of The Curator's Machine is based on ImageNet-trained weighted networks. A bias in favour of texture and facture could result in classifications that are unusual and new for human viewers oriented towards shape, making other insights possible. Users should therefore be made aware of the ways in which texture bias may take effect. They might even be provided intentionally with a switch between texture and shape based methods.

Fig. 11: Lucas Cranach the Younger, Martin Luther (1548): When computer vision detects this woodcut, the age spots or the structure of the edge of the paper can create classifications or k-nearest neighbors through texture bias that are undesirable from a human perspective (detail, woodcut, Statens Museum for Kunst, Copenhagen, public domain, KKSgb5082). The figure serves to illustrate the problem according to Geirhos et al. 2020. No individual verification for the prototype The Curator's Machine was carried out based on the figure.

Fig. 12: Lucas Cranach the Elder. Martin Luther (1532): detail with attention to craquelure (detail, oil painting, Statens Museum for Kunst, Copenhagen, public domain, KMSsp720). See previous figure note. In addition, it remains to be explored to what extent the fine-grained craquelure is still relevant at 244×244 pixels. The detection of craquelure has earlier been discussed in Description and Classification of Craquelure (Bucklow 1999) and was recently applied to weighted networks in Craquelure as a Graph: Application of Image Processing and Graph Neural Networks to the Description of Fracture Patterns (Sidorov and Hardeberg 2019).

6. Conclusion

This paper explores issues and pitfalls with ImageNet trained weighted neural networks which may influence the software prototype The Curator's Machine. By examining the non-classification of 'art' in ImageNet, the even more concerning ahistoricity of the classifications, and the tendency of ImageNet trained weighted networks to favor texture over shape, a number of conclusions can be drawn:

1. As far as *feature extraction* is concerned, tests should be carried out to see whether the next prototype of The Curator's Machine can be developed with weighted networks that are not based solely on ImageNet. The extent to which the feature extraction of the pre-trained networks affects the resulting modules of The Curator's Machine should be investigated. Other image data collections that show better recognition performance in relation to 'art', such as Google's Open Images Dataset, should be examined for their suitability.
2. The lack of historicity in the existing training datasets is a semantic barrier that can only be overcome, if at all, by metadata trained into the computer vision networks. Connecting Text and Images (CLIP) networks can be used for this purpose, but they introduce further complexity.
3. Classic (ImageNet) pre-trained networks favour texture. The users of The Curator's Machine should be made aware of this fact. Ideally, the difference between texture and shape should be included as a selectable option.

Open Research Questions

Transfer learning using Open Images and other, smaller datasets can investigate the extent to which certain art-historically relevant features of two-dimensional works can be learned. Computer vision can only to a limited extent process multidimensional art, including time-based, ephemeral, or conceptual strategies. When these 'complicated' media get excluded, the exclusions should be marked more clearly for the user.

Thinking of 'the user,' apart from the researchers and developers of a weighted networks, in a weighted networks based software application, shifts the perspective towards questions of human-computer interaction and user interface design. This includes a perspective, where weighted networks are not stand-alone developments, but embedded into layers of software and software infrastructure. Therefore user-interface related questions result from this paper, for instance how to present the difference between texture and shape bias, or how to visually demonstrate the ahistoricity of training data and its effects on operative images.

To be able to reformulate the user interface, more research is needed to explore to what extent the effects of the ahistoricity of ImageNet described here are actually expressed in the grids, scatter plots, and other visual classification diagrams. This would be a task for a specialised study.

When pre-trained weighted networks get used not for classification tasks but for feature extraction and subsequent tasks of cultural analytics, it needs to be investigated, whether the already learned features in partial weighted networks are sufficient for the task. In other words: Can domain-specific projects like The Curator's Machine build on non-domain-specific training sets for feature extraction? Looking into these issues might lead to adopting other general approaches, for instance the usage of Autoencoders or of Contrastive Language–Image Pre-training (CLIP) techniques.[24]

NOTES

1 To de-anthropomorphise the discourse, the term 'weighted networks' is used here instead of 'artificial neural networks.' These networks, which were originally conceived in the 1960s as being similar to neurons (Rosenblatt 1957), are characterised by weighted nodes, which do not correspond to the function of neurons in the human body according to current scientific knowledge (Cardon, Cointet, and Mazieres 2018, 8).

2 https://github.com/DominikBoenisch/Training-the-Archive/.

3 The use of the term computer vision can be traced back to the 1960s. In contrast to digital image processing, which refers to the automated processing of two-dimensional images and addresses issues such as character recognition (OCR), computer vision should be able to automate complex image relationships, such as detecting the movements and interactions of objects in images. This approach is ultimately aimed at decision-making systems.

4 Recommended introductory resources from a humanities perspective include: *How the machine 'thinks' – Understanding Opacity in Machine Learning Algorithms* (Burrell 2016, 5–7) and *How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence* (Pasquinelli 2019, 4–14).

5 On the genealogy of ImageNet, see also *Excavating AI* (Crawford and Paglen 2019) and *Lines of Sight* (Hanna et al. 2020).

6 The procedure is presented here in simplified form. In reality, the features of various pre-trained networks (InceptionV3, BiT/m-r152×4 and individual layers from VGG19) were extracted for each of the 42,000 input images from the SMK collection and then further processed. All these networks are pre-trained with ImageNet.

7 See also: https://github.com/DominikBoenisch/Training-the-Archive/blob/master/Prototype/2_Feature_Extractor/Feature_Extractor_Keras_Applications.ipynb.

8 Since the weighted networks for the prototype were initialized using Keras/Tensorflow, the implementation there is decisive. Compare: https://keras.io/api/applications/.

9 The optimal pixel values are determined experimentally and have meanwhile become established as a standard. One of the reasons seems to be that a weighted network's smallest feature map should be able to process the image and the input data should not be too large: "In general, initialized networks with an input size of 224×224px obtained the best results: Xception, InceptionV3, and MobileNet (above 99%); SqueezeNet also obtained competitive results (98.36%) with a smaller input (192×192px)" (Alashhab, Gallego, and Lozano 2019, 4).

10 When building image recognition applications, programmers can either train a weighted network from the scratch or use ready-made 'pre-trained' networks, which have been published by third parties. ImageNet is an example for such a pre-trained network. The training process involves feeding labeled images through the network and automatically fine-tuning the network using 'back-propagation.' Basically, the 'trainer' feeds 1000 images with cats and the label 'cat' attached through the network, until the weights within the network are sufficiently tuned. An imprinted recognition then can be transferred to images without labels, so that 'unknown' images can be detected by a pre-trained weighted network.

11 An overview of the weighted networks used on Wolfram Alpha is available at https://resources.wolframcloud.com/NeuralNetRepository. The image recognition function used by Lee is based on a specially developed weighted network Wolfram ImageIdentify Net V1. The manufacturer does not provide any information about the training data, even upon request.

12 In Wordnet, a search for 'art' yields four different synsets (ID: 02746552, 00935235, 05646832, 07011408), of which 'artwork' (ID: 07011408) is the most relevant for our purposes (cf. http://wordnetweb.princeton.edu/perl/webwn?s=art&sub=Search+WordNet, retrieved 11/03/2021). The noun 'painting' (ID: 03882197, 00938436) is also represented by four synsets, although two refer to painting as a craft, not painting as a work of art (ID: 00718460, 00610504). Since some paintings were recognised as art if they had a frame, according to Lee, 'frame' in the sense of picture frame (ID: 03395829) was also searched for.

13 About the procedure: The original ImageNet Fall 2011 list contained the existing synset categories by ID number (http://image-net.org/ImageNet_data/urls/ImageNet_fall11_urls.tgz, accessed on 13/03/2017 – this link no longer works). These were the URLs from which training data could be downloaded. To a large extent, this contains images from Flickr, some of which are under Creative Commons licence, and some of which are unlicensed. The IDs we were looking for could not be found in this list. ImageNet LSVRC 2012 contains 1000 classes with 1.28 million images, but also none of the classes searched for, see https://raw.githubusercontent.com/mf1024/ImageNet-Datasets-Downloader/master/classes_in_imagenet.csv (accessed on 11/03/2021). The same applies to the ImageNet 21K collection with 21,000 image classes, see https://github.com/dmlc/mxnet-model-gallery/blob/master/imagenet-21k-inception.md (accessed on 11/03/2021).

14 This is where things become complex, because the VGG19 modules, which were also trained via ImageNet, were introduced with a lower weighting (90%), according to the principle of style transfer, i.e., they detect those textures that computer scientists refer to as "styles." These image styles are problematic in themselves, as they are not based on art-historical expertise, but rather follow a popular understanding of art, yet appear here in the guise of science (cf. Gatys, Ecker, and Bethge 2015, 6).

15 See https://microscope.openai.com (accessed on 11/03/2021).

16 In the following layers of the network, a similarly strong reference to drapery could no longer be found.

17 Beyond this observation, the text by Goh et al. is problematic because it attempts to map certain facial features onto human emotions, and to identify this in turn in the 'neurons' of weighted networks. Such phrenological illusions are opposed by Bowyer et al. in *The "Criminality from Face" Illusion* (Bowyer et al. 2020), Stinson in *The Dark Past of Algorithms That Associate Appearance and Criminality* (Stinson 2020) and Munn in *Logic of Feeling – Technology's Quest to capitalize Emotion* (Munn 2020).

18 Continuation of the quote: "This is particularly disappointing because a popular approach in computer vision is to use convolutional network features as a space where Euclidean distance approximates perceptual distance. This resemblance is clearly flawed if images that have an immeasurably small perceptual distance correspond to completely different classes in the network's representation" (ibid.).

19 See: https://artsandculture.google.com/explore (accessed on 15/03/2021).

20 Piet Mondrian's modernist composition, *Composition en blanc et noir II* from 1930, is, for instance, detected by Google Cloud Vision as "product design, 68%; furniture 66%, picture frame 52%, window 51%" and On Kawara's *JULY 4, 1973 Wednesday* as "font 78%, product design 62%, brand 62%."

21 In this context, ageographic means the indiscriminate mixing of the most diverse visual cultures in the training sets, as exemplified by the Open Images Dataset, which treats Asian and European sculpture as sculpture and suppresses the cultural genesis of the respective artistic techniques and subjects.

22 See https://github.com/rgeirhos/Stylized-ImageNet (accessed on 15/03/2021). Pre-trained stylised weighted networks are provided by Geirhos et al. at https://github.com/rgeirhos/texture-vs-shape.

23 Since they suggest the procedure for general image sets, it is worth mentioning which data collection they used for the style transfer: Kaggle's *Painter by Numbers*, which is largely based on painting images collected on WikiArt. See https://www.kaggle.com/c/painter-by-numbers/ und https://www.wikiart.org (accessed on 15/03/2021).

24 The author gratefully acknowledges comments and contributions by Inke Arns, Dominik Bönisch, Matthias Pitscher, Nicolas Malevé und Alexa Steinbrück.

BIBLIOGRAPHY

Alashhab, Samer, Antonio-Javier Gallego, and Miguel Ángel Lozano. 2019. "Hand Gesture Detection with Convolutional Neural Networks." In *Distributed Computing and Artificial Intelligence, 15th International Conference*, edited by Fernando De La Prieta, Sigeru Omatu, and Antonio Fernández-Caballero, 45–52. Cham: Springer International Publishing.

Amat, Josep, and Alicia Casals. 1992. "Image Obtention and Preprocessing." In *Computer Vision: Theory and Industrial Applications*, edited by Carme Torras, 1–58. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bönisch, Dominik. 2021. "Curator's Machine – Clustering Museum Collection Data by Annotating Hidden Patterns of Relationships between Artworks." *International Journal for Digital Art History*, no. 5 (May): 20–35. doi:10.11588/DAH.2020.5.75953.

Bowyer, Kevin W., Michael C. King, Walter J. Scheirer, and Kushal Vangara. 2020. "The 'Criminality From Face' Illusion." *IEEE Transactions on Technology and Society* 1 (4): 175–83. doi:10.1109/TTS.2020.3032321

Broeckmann, Andreas. 2016. *Machine Art in the Twentieth Century*. Leonardo Book Series. Cambridge, MA: MIT Press.

Brownlee, Jason. 2019. "Best Practices for Preparing and Augmenting Image Data for CNNs." Blog. *Machine Learning Mastery*. May 2. https://machinelearningmastery.com/best-practices-for-preparing-and-augmenting-image-data-for-convolutional-neural-networks/.

Burrell, Jenna. 2016. "How the Machine 'Thinks' – Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1). doi:10.1177/2053951715622512.

Cardon, Dominique, Jean-Philippe Cointet, and Antoine Mazieres. 2018. "Neurons Spike Back: The Invention of Inductive Machines and the Artificial Intelligence Controversy." *Reseaux* 36 (211): 173–220. doi:10.3917/res.211.0173.

Chaki, Jyotismita, and Nilanjan Dey. 2020. *A Beginner's Guide to Image Preprocessing Techniques*. Boca Raton London: CRC PRESS.

Crawford, Kate, and Trevor Paglen. 2019. "Excavating AI – The Politics of Images in Machine Learning Training Sets." Website. *Excavating AI*. September 19. https://www.excavating.ai.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. 2015. "A Neural Algorithm of Artistic Style." *ArXiv:1508.06576* [Cs, q-Bio], September. http://arxiv.org/abs/1508.06576.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. "Shortcut Learning in Deep Neural Networks." *ArXiv:2004.07780* [Cs, q-Bio], May. http://arxiv.org/abs/2004.07780.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. "ImageNet-Trained CNNs Are Biased towards Texture – Increasing Shape Bias Improves Accuracy and Robustness." *ArXiv:1811.12231*, January. http://arxiv.org/abs/1811.12231.

Goh, Gabriel, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. "Multimodal Neurons in Artificial Neural Networks." *Distill* 6 (3). doi:10.23915/distill.00030.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. "Explaining and Harnessing Adversarial Examples." *ArXiv:1412.6572*, March. http://arxiv.org/abs/1412.6572.

Groys, Boris. 2004. *Über Das Neue – Versuch Einer Kulturökonomie*. 3rd ed. Fischer Forum Wissenschaft Kultur & Medien. Frankfurt am Main: Fischer-Taschenbuch-Verl.

Hanna, Alex, Emily Denton, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2020. "Lines of Sight." Online Magazin. *Logic Magazine*. December. https://logicmag.io/commons/lines-of-sight/.

Hinton, Geoffrey E., Alex Krizhevsky, and Ilya Sutskever. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 1097–1105. NIPS'12. USA: Curran Associates Inc.

Huang, Xun, and Serge Belongie. 2017. "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization." *ArXiv:1703.06868*, July. http://arxiv.org/abs/1703.06868.

Huh, Minyoung, Pulkit Agrawal, and Alexei A. Efros. 2016. "What Makes ImageNet Good for Transfer Learning?" *ArXiv:1608.08614* [Cs], December. http://arxiv.org/abs/1608.08614.

Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. 2019. "Do Better ImageNet Models Transfer Better?" *ArXiv:1805.08974* [Cs, Stat], June. http://arxiv.org/abs/1805.08974.

Kunsthistorisches Institut der CAU Kiel. 2019. "Fachausdrücke zur Benennung Und Beschreibung von Gewandfalten Figürlicher Darstellungen–Bildkünste." CAU Kiel. https://www.kunstgeschichte.uni-kiel.de/de/infos-fuer-das-studium/fachausdrucke-bildkunste-ss-2019-neu.pdf.

Lee, Rosemary. 2020. "Machine Learning and Notions of the Image." Dissertation, Copenhagen: Center for Computer Games Research, Department of Digital Design, IT-University of Copenhagen. https://en.itu.dk/~/media/en/research/phd-programme/phd-defences/2020/phd-thesis-final-version-rosemary-lee-pdf.pdf?la=en.

Li, Fei-Fei, Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, and Kai Li. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. Miami, FL: IEEE. doi:10.1109/CVPR.2009.5206848.

Malevé, Nicolas. 2020. "On the Data Set's Ruins." *AI & SOCIETY*. doi:10.1007/s00146-020-01093-w.

Munn, Luke. 2020. *Logic of Feeling – Technology's Quest to Capitalize Emotion*. Lanham: Rowman & Littlefield.

Offert, Fabian, and Peter Bell. 2020. "Perceptual Bias and Technical Metapictures – Critical Machine Vision as a Humanities Challenge." *AI & SOCIETY*, October. doi:10.1007/s00146-020-01058-z.

Parikka, Jussi. 2021. "On Seeing Where There's Nothing to See – Practices of Light beyond Photography." In *Photography off the Scale – Technologies and Theories of the Mass Image*, edited by Jussi Parikka, 185–210. Edinburgh: Edinburgh University Press.

Pasquinelli, Matteo. 2019. "How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence." *Spheres. Journal for Digital Cultures.*, no. 5 (November): 1–17.

Pereira, Gabriel, and Bruno Moreschi. 2020. "Artificial Intelligence and Institutional Critique 2.0 – Unexpected Ways of Seeing with Computer Vision." *AI & SOCIETY*, September. doi:10.1007/s00146-020-01059-y.

Rosenblatt, Frank. 1957. "The Perceptron – A Perceiving and Recognizing Automation." 85-460–1. Buffalo, NY: Cornell Aeronautical Laboratory.

Sauerländer, Willibald. 1970. *Gothic Sculpture in France*, 1140-1270. New York, NY: Harry N. Abrams.

Schmitt, Philipp. 2019. "Declassifier." Website. *Humans of AI*. 2020. https://humans-of.ai/.

Sidorov, Oleksii, and Jon Yngve Hardeberg. 2019. “Craquelure as a Graph: Application of Image Processing and Graph Neural Networks to the Description of Fracture Patterns.” In *2019 IEEE/CVF International Conference on Computer Vision Workshop* (ICCVW), 1429–36. Seoul, Korea (South): IEEE. doi:10/gm5v77.

Stack, John. 2019. “What the Machine Saw.” Website. *What the Machine Saw*. https://johnstack.github.io/what-the-machine-saw/index.html.

Stinson, Catherine. 2020. “The Dark Past of Algorithms That Associate Appearance and Criminality – Machine Learning That Links Personality and Physical Traits Warrants Critical Review.” Online Publication. *American Scientist*. December 2. https://www.americanscientist.org/article/the-dark-past-of-algorithms-that-associate-appearance-and-criminality.

Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. “How Transferable Are Features in Deep Neural Networks?”ArXiv:1411.1792 [Cs], November. http://arxiv.org/abs/1411.1792.

FRANCIS HUNGER (Ph.D.). His practice combines artistic research and media theory with the capabilities of narration through installations, radio plays and performances and internet-based art. Currently he is a researcher for the project Training The Archive at Hartware MedienKunstVerein, Dortmund, critically examining the use of AI, statistics and pattern recognition for art and curating. In 2022 he co-curated with Inke Arns and Marie Lechner the exhibition House of Mirrors – Artificial Intelligence as Phantasm at HMKV, Dortmund. His Ph.D. at Bauhaus University Weimar developed a media archeological genealogy of database technology and practices. In 2022/23 Hunger was guest professor at the Intermedia program of the Hungarian Academy for Visual Arts, Budapest. Recent texts include Data Workers of All Countries, End It! Hamburg 2022, Transaktionsverarbeitung in relationalen Datenbanken – Zur Materialität von Daten aus Perspektive der Transaktion. Paderborn 2021, How to Hack Artificial Intelligence. Artistic projects and current research on the (dis) abilities of machine learning techniques. Amsterdam 2019, and Epistemic Harvest – The electronic database as discourse and means of data production. Aarhus 2018 Hunger’s artistic work is exhibited internationally. Numerous festival participations, talks, lectures, publications, screenings and academic lectures. He occasionally curates exhibitions, teaches at universities regularly and publishes daily on twitter. http://www.irmielin.org, http://twitter.com/databaseculture

Correspondence e-mail: francis.hunger@irmielin.org.